PROGNOSTIC MODELS OF CLINICAL OUTCOMES AND PREDICTIVE MODELS OF TREATMENT RESPONSE IN PRECISION PSYCHIATRY

PROGNOSTIC MODELS OF CLINICAL OUTCOMES AND PREDICTIVE MODELS OF TREATMENT RESPONSE IN PRECISION PSYCHIATRY

By Devon Patrick Watts, M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

McMaster University DOCTOR OF PHILOSOPHY (2022) Hamilton, Ontario (Neuroscience)

TITLE: Prognostic Models of Clinical Outcomes and Predictive Models of Treatment Response in Precision Psychiatry

AUTHOR: Devon Patrick Watts

M.Sc. (McMaster University).

SUPERVISOR: Dr. Flavio Kapczinski, M.D., M.Sc., Ph.D.

COMMITTEE: Dr. James P. Reilly, Ph.D.; Dr. Kathryn Murphy, Ph.D.

NUMBER OF PAGES: xxiii, 374.

Abstract

In this thesis, we developed prognostic models of clinical outcomes, specific to violent and criminal outcomes in psychiatry, and predictive models of treatment response at an individual level. Overall, we demonstrate that evidence-based risk factors, protective factors, and treatment status variables were able to prognosticate prospective physical aggression at an individual level; 2) prognostic models of clinical and violent outcomes in psychiatry have largely focused on clinical and sociodemographic variables, show similar performance between identifying true positives and true negatives, although the error rate of models are still high, and further refinement is needed; 3) within treatment response prediction models in MDD using EEG, greater performance was observed in predicting response to rTMS, relative to antidepressants, and across models, greater sensitivity (true positives), were observed relative to specificity (true negatives), suggesting that EEG prediction models thus far better identify non-responders than responders; and 4) across randomized clinical trials using data-driven biomarkers in predictive models, based on the consistency of performance across models with large sample sizes, the highest degree of evidence was in predicting response to sertraline and citalopram using fMRI features.

Keywords: precision psychiatry; computational neuroscience; psychotic disorders; genomics

Acknowledgements

One of the great Roman emperors, Marcus Aurelius, once famously wrote that, "…the impediment to action advances action. What stands in the way becomes the way". I can't seem to think of a more perfect metaphor for my experience throughout graduate school. Between unforeseen roadblocks, bureaucratic constraints, and the necessity of pivoting away from conducting new experiments due to university COVID-19 policies, these experiences provided a series of growth-opportunities that I'm ultimately quite grateful for. Without them, it is unlikely that I would have focused on computational neuroscience, where I was provided the geographic freedom to work anywhere in the world, provided I had a stable internet connection and access to a computer.

More importantly, however, I'm very fortunate for supportive friends, colleagues, and my academic supervisor. I'd like to first thank my mentor and friend Dr. Flavio Kapczinski for affording me the high degree of latitude to explore a research focus that I found both meaningful and impactful. I will always be grateful for the lessons you have taught me in navigating academia and charting your own path as an independent scientist. I would also like to thank Taiane Cardoso for her continual support, friendship, and important insights throughout my graduate school experience. You have been an exemplar for how to be a successful young scientist, and how to foster a positive and collaborative environment. I would also like to thank my collaborators, including Ives Cavalcante Passos, Gary Chaimowitz, Mini Mamak, and Heather Moulden. Your insights have been invaluable throughout my PhD, and I owe a great deal of gratitude for fostering my interest in developing predictive machine learning models with real-world clinical utility. I would also like to thank Florence Roullet for all her help and

patience with the logistics of traveling throughout the last few months to facilitate collaborations and navigate the complicated and opaque process of utilizing research funds.

I would also like to thank Daniela Russo for her continual support, patience, and willingness to help problem solve with me, especially during times where everything was metaphorically on fire. Words cannot express my gratitude, and I will be sure to return the favour when you experience the rollercoaster of graduate school. I'd also like to personally thank her parents, Annamaria and "Big Dave" Russo for welcoming me into their home, their wisdom, and hospitality. My transition to the US wouldn't have been possible without the hectic semi-nomadic lifestyle I lived last year travelling constantly back and forth between Hamilton and Boston and largely living out of a suitcase. I'd also like to thank my Mom, Tina Watts, and Gerhard "Hart" Proksch for their continual support, and always being in my corner.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AdaBoost	Adaptive Boosting
AIS	Aggressive Incidents Scale

AUD	Alcohol-Use Disorder
Apeglm	Approximate Posterior Estimation
AA	Arachidonic acid
AUC	Area-Under-the-Curve
BAM	Binary Alignment Map
BPRS	Brief Psychiatric Rating Scale
CATIE	Clinical Antipsychotic Trials of Intervention Effectiveness
CAPS	Clinician-Administered PTSD Scale
CBT	Cognitive Behavioural Therapy
CPM	Connectome-Based Predictive Modelling
DOR	Diagnostic odds ratio
DHA	Docosahexaenoic acid
DLPFC	Dorsolateral Prefrontal Cortex
DTI	Diffusion Tensor Imaging
EEG	Electroencephalography
e-HARM	Electronic Hamilton Anatomy of Risk Management
EPA	Eicosapentaenoic acid
ERG	Electroretinogram
EMBARC	Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care for Depression
EBM	Evidence-based medicine
f-idf	frequency-inverse document frequency
fNIRS	functional near-infrared spectroscopy
GLM	Generalized linear model
GWAS	Genome-wide association study
GAF	Global Assessment of Functioning
GBM	Gradient boosting machine
HRF	Haemodynamic response function

HAM-D	Hamilton Depression Rating Scale
HCR-20	Historical, Clinical and Risk Management
IAF	Individual alpha frequency
ISPC	International SSRI Pharmacogenomics Consortium
ICN	Intrinsic connectivity network
IDS-30	Inventory of depressive symptomatology—self-rated
k-NN	k-Nearest Neighbours
MDD	Major Depressive Disorder
MEG	Magnetoencephalography
MICE	Multiple imputation by chained equations
MADRS	Montgomery-Asberg Depression Rating Scale
MRI	Magnetic Resonance Imaging
MARS	Multivariate adaptive regression splines
negLR	Negative likelihood ratio
NRS-PM	Network restricted strength predictive model
NBS	Network-based statistics
neNRS	Nodal external network restricted strength
niRNS	Nodal internal network restricted strength
NCR	Not criminally responsible
NAP1L4	Nucleosome Assembly Protein 1 Like 4
NPV	Negative predictive value
ω-3	Omega-3 fatty acids
PGNG	Parametric go/no-go test
POSTN	Periostin
PGRN-AMPS	Pharmacogenomic Research Network Antidepressant Medication Pharmacogenetic Study
PUFAs	Polyunsaturated fatty acids
posLR	Positive likelihood ratio

PMI	Post-mortem interval
PTSD	Post-Traumatic Stress Disorder
PPV	Positive predictive value
PCA	Principal component analysis
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies-2
QIDS	Quick Inventory of Depressive Symptomatology
RCT	Randomized controlled trial
ROC	Receiver operating characteristic
ROIs	Regions of interest
RVM	Relevance vector machine
SCZ	Schizophrenia
SHAP	SHapley Additive exPlanations
SNP	Single-nucleotide polymorphism
SELSER	Sparse EEG Latent SpacE Regression
SMRI	Stanley Medical Research Institute
STAR*D	Sequenced Treatment Alternatives to Relieve Depression
SVM	Support Vector Machine
SVA	Surrogate variable analysis
SMOTE	Synthetic minority oversampling technique
$\tau 2$	Tau squared
tDCS	Transcortical direct current stimulation
TMS	Transcranial magnetic stimulation
TWAS	Transcriptome-wide association study
UST	Unfit to stand trial
VRAG	Violence Risk Appraisal Guide-Revised
XGBoost	Extreme Gradient Boosting

Declaration of Academic Achievement

Chapter 2

D. P. Watts: Conceptualization, Data curation, Statistical Analyses, Methodology, Original draft, Final draft. M. Mamak: Conceptualization, Final draft. H. Moulden: Conceptualization, Final draft. C. Upfold: Data curation, Final draft. T.A. Cardoso: Original draft, final draft. Kapczinski F: Original draft, Final draft. G. Chaimowitz: Funding acquisition, project administration, conceptualization, Final draft.

This chapter in its entirety has been ***submitted*** to the **Journal of Psychiatric Research**.

Chapter 3

D. P. Watts: Conceptualization, Screening papers, conducting the meta-analysis, initial and final draft, interpreting the findings, creating figures. D. Librenza-Garcia: Conceptualization, Screening papers. P. Ballester: Screening papers and interpreting the findings. T.A. Cardoso, I.C. Passos, F.H.P. Kessler, J. Reilly, F. Kapczinski, and G. Chaimowitz supervised the work and interpreted the findings.

This chapter in its entirety has been ***accepted*** to the journal **Translational Psychiatry.**

Chapter 4

All authors participated in the writing, revisions, and the approval of the final manuscript.

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

This chapter in its entirety has been ***published*** in the journal **Trends Psychiatry Psychotherapy**. The final accepted manuscript version of this article is presented within this thesis.

Watts D, D'Souza J, Azevedo MA, Kapczinski F, Chaimowitz G. Stigmatized individuals: a case for precision ethics. Trends Psychiatry Psychother. 2021 Sep 2. doi: 10.47626/2237-6089-2021-0354. Epub ahead of print. PMID: 34551242.**Chapter 5**

All authors participated in the writing, revisions, and the approval of the final manuscript.

This chapter in its entirety has been ***published*** in the Brazilian Journal of Psychiatry. The final accepted manuscript version of this article is presented within this thesis.

Watts D, Garcia FD, Lacerda ALT, Mari JJ, Quarantini LC, Kapczinski F. Intranasal esketamine and the dawn of precision psychiatry. Braz J Psychiatry. 2022 Mar-Abr;44(2):117-118. doi: 10.1590/1516-4446-2021-0031. PMID: 34320126; PMCID: PMC9041972.

Chapter 6

D.P. Watts: Conceptualization, Methodology, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization. R.F. Pulice: Methodology, Writing – Original Draft. J Reilly: Writing – Review & Editing. A. Brunoni: Writing – Review & Editing. F. Kapczinski: Writing – Review & Editing. Ives Cavalcante Passos: Conceptualization, Methodology, Writing – Review & Editing.

This chapter in its entirety has been ***published*** in the journal **Translational Psychiatry**.

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

Watts, D., Pulice, R.F., Reilly, J. *et al.* Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Transl Psychiatry* **12**, 332 (2022). https://doi.org/10.1038/s41398-022-02064-z

Chapter 7

D. P. Watts: Conceptualization, Screening papers, initial and final draft, visualization, and interpreting the findings. D. Librenza-Garcia: Conceptualization, Screening papers. P. Ballester: Screening papers, interpreting findings, and visualization. B.J. Kotzian: initial draft. J. Yang: Initial draft. B. Frey, L. Minuzzi, B. Mwangi, F. Kapczinski, and I.C. Passos, supervised the work, interpreted the findings, and participated in the final draft.

This chapter in its entirety is currently ***under revision*** in the journal **Molecular Psychiatry**.

Chapter 1: Introduction

1.1. Overall Approach: Precision Psychiatry

Precision psychiatry is an emerging field that seeks to advance the personalized care of patients through improving our capacity to prognosticate the probability of clinical outcomes (prognostic models), response to therapeutic interventions (predictive models), and identify the presence of specific disorders (diagnostic models) at an individual patient level [1]. Recent developments in individualized predictive modeling and large-scale data collection in psychiatry have facilitated a renewed effort to address longstanding issues with determining individual patient risk of clinical outcomes, a trial-and-error approach to treatment, as well as identifying reliable and valid biological diagnostic markers of either specific disorders, or biological markers that underlie systems across disorders, such as negative and positive valence in the context of the NIMH Research Domain Criteria (RDoC) [2,3].

This thesis contributes to toward the field of precision psychiatry through developing prognostic and predictive models of negative clinical outcomes and treatment response, respectively. Principally, we develop prognostic models of prospective physical aggression in patients with Schizophrenia, where the performance of data-driven models is compared to clinician-rated clinical judgement of immediate and short-term violent risk, to assess its utility relative to standard clinical practices. We also developed a meta-analysis of predicting violent and criminal outcomes in psychiatric patients, to identify important features, as well as evaluate the predictive capabilities across clinical models. In an editorial, we also consider the ethical and legal ramifications of prognostic models of criminality and violence in psychiatry and offer considerations to minimize patient harm. With respect to predictive models in precision psychiatry, we first consider the use case of ketamine in treatment-resistant depression and highlight the need for identifying candidate

biological markers that can predict treatment response at an individual level. Furthermore, we compiled the evidence-base for predicting treatment response using electroencephalography (EEG), a cost-effective neurophysiological measure of brain activity with excellent temporal resolution, in the context of Major Depressive Disorder (MDD), which comprises most EEG treatment response prediction models in psychiatry to date. Moreover, we synthesize existing literature on data-driven biomarkers of treatment response within clinical trials in psychiatry and introduce a concept of precision machine learning trials as a candidate trial design.

1.2. Predicting clinical outcomes in schizophrenia

1.2.1. Epidemiology

While schizophrenia is a relatively rare disorder, with a global age-standardized point prevalence of 0.28% (95% uncertainty interval: 0.24-0.31) across 195 countries and territories, it is associated with a substantial burden of disease [4]. Similarly, in a systematic review comprising studies from 47 countries, with an estimated 154,140 potentially overlapping cases of schizophrenia, the median point prevalence of schizophrenia was 4.6 per 1000 lives, which measures prevalence at a particular point in time, the period prevalence was 3.3 per 1000 lives, which is the proportion of individuals with a disease or attribute at any time during the interval, and the lifetime prevalence was 4.0 per 1000 lives, respectively [5].

Although heterogeneity exists across patients, the core characteristics of schizophrenia include positive symptoms, negative symptoms, and cognitive symptoms. Positive symptoms involve the presence hallucinations, or sensory experiences in the absence of external stimuli, and delusions, that involve bizarre or irrational beliefs that are incongruent with broader society [6]. Conversely, negative symptoms comprise the decrease or absence of normal behaviors and functions [7],

including affective flattening, decreased movements, lack of vocal inflection, poverty of speech and content, avolition/apathy, anhedonia/asocial behavior, and emotional withdrawal [8]. Moreover, impairments in executive functioning, while present in other forms of psychosis [9], tends to be more severe, with an earlier onset, and independent of other clinical symptoms in the context of Schizophrenia [10,11].

1.2.2. Poor clinical outcomes are common in schizophrenia

Apart from the core symptomatology of schizophrenia, there has been a large body of literature indicating poor clinical outcomes among these patients. For instance, only one in seven patients are expected to meet criteria for long-term recovery [12], and in a 15-year prospective follow-up study, less than 40% of patients with schizophrenia showed one or more periods of recovery [13]. Several risk factors of poor clinical outcomes have been identified across literature, including duration of the disorder, severity, cognitive impairment, and insight into illness.

With respect to disorder duration, in a systematic review and meta-analysis across 33 studies, a small statistically significant correlation coefficient (r =0.13-0.18) was observed between long duration of untreated psychosis and poor general symptomatic outcomes, greater overall symptoms, decreased likelihood of remission, as well as poor social functioning [14]. Additionally, within a separate meta-analysis across 43 studies, greater severity of negative symptoms was found to be significantly associated with a longer untreated psychosis duration (combined Hedges's gu=0.28, 95% CI=0.1–0.45; combined correlation: r=0.15, 95% CI=0.9–0.21). Moreover, untreated psychosis duration was not found to be related to positive symptoms, global assessments of functioning, or global psychopathology severity at initial treatment contact [15]. In another meta-analysis of 4490 participants, across 26 studies, while no statistically significant differences were observed between short and long duration on any outcome at 24 months,

including symptom severity, symptom domain, quality of life, and social functioning, at 15-year follow-up, patients with an untreated psychosis duration showed significantly worse outcomes in all domains, apart from negative symptoms, and were less likely to achieve remission. An association was also observed between longer duration of untreated psychosis and worse outcomes at 6 months, including overall functioning, quality of life, and total symptoms [16].

Furthermore, among studies examining the role of symptom severity, within a prospective follow-up of previously treated (N=45) and first-episode patients with schizophrenia (N=53) assessed at intake with functional outcomes evaluated 2-8 years later (average 3 years), a higher level of overall functioning at follow-up was predicted by lower levels of depressive, positive, and negative symptoms at intake [17]. Additionally, within a recent 3-year follow-up longitudinal retrospective study, patients with high levels of primary negative symptoms, which are characterized as largely persistent across illness stage and overall lifespan, individuals with higher levels of primary negative symptoms showed worse psychosocial functioning, poorer cognitive performance, earlier age of disorder onset, and greater utilization of psychiatric services, including a higher number of admissions in acute care, and overall inpatient services [18].

Prior studies have also investigated the impact of insight into illness in schizophrenia, which involves whether a patient recognizes that they possess an illness that requires treatment or remediation. In a sample of 96 patients with a diagnosis of schizophrenia, 58.2% of patients lacked insight into symptoms, 32.7% for illness, 18.4% for treatment response, and 41.8% lacked understanding of the social consequences of their disorder [19]. Poor insight into illness has also been associated with treatment compliance and psychosocial functioning [20] among patients. However, available evidence suggests that improving insight into illness is challenging among

patients with schizophrenia. For example, in a randomized controlled trial of compliance therapy in 56 individuals with schizophrenia, no major advantage was found over non-specific therapy in improving compliance at one year, or in any secondary outcome measures, including insight, global assessment of functioning, quality of life or overall symptomatology [21]. Furthermore, although inconsistencies exist across the literature, lack of insight has been suggested to be a risk factor for aggressive behavior in schizophrenia [22]. In a sample of 115 violent patients with schizophrenia within forensic settings, and 111 patients with schizophrenia without a history of violent behavior, violent patients showed poorer functioning, less insight, and were more symptomatic [22]. Furthermore, in a sample of 47 patients with violent schizophrenia and 86 nonviolent patients, those without a history of violence showed lower positive symptom scores, and higher clinical insight. Moreover, delusional severity, history of violence, and worse clinical insight were found to be significant predictors of violence in the context of schizophrenia [23]. Altogether, patients with schizophrenia tend to show poor rates of remediation, and there is growing evidence of negative behavioral outcomes in a subset of individuals including violence and criminality.

1.2.3. Criminal and violent outcomes in schizophrenia

Indeed, across prior studies, there is evidence to suggest that schizophrenia and related disorders are associated with increased rates of violent crime [24] and violent risk, particularly in those with symptoms of delusional beliefs [25]. Considering this, the prediction and prevention of aggression in patients with psychotic disorders remains among the top priorities in their clinical care [26]. In a recent meta-analysis and systematic review of the association between schizophrenia spectrum disorders and the perpetration of violence comprising 51,309 individuals across 24 studies in 15 countries over four decades [27], those with psychosis and comorbid substance misuse showed

approximately 10-fold increased odds of prospective violence relative to general population controls. However, this relative risk was much lower, approximately 3-fold, among individuals lacking comorbidities [27]. Similarly, another meta-analysis comprising 204 studies across 166 independent datasets suggests that psychosis is associated with a 49-68% increased likelihood of violence, although substantial variability was found due to moderating factors such as how psychosis is measured, the presence of a comparison group, and study design [27]. Altogether, this highlights the necessity of identifying risk factors and applying preventative strategies among a subset of patients who show an elevated likelihood of prospective physical aggression.

1.2.4. Risk factors of criminality in schizophrenia spectrum disorders

Within prior studies, several modifiable and causal risk factors for violent outcomes in psychosis have been identified including treatment nonadherence [28], impulsivity [29], and childhood trauma [30]. For example, in a prospective longitudinal UK Prisoner Cohort Study [31], comprising individuals without psychosis (N=742), with schizophrenia (N=94), delusional disorder (N=29), and drug-induced psychosis (N=102), schizophrenia was found to be associated with violence, but only in the absence of treatment (Odds Ratio, OR = 3.76, 95% CI: 1.39-10.19). Moreover, untreated schizophrenia was associated with the appearance of persecutory delusions at follow-up (OR = 3.52, 95% CI: 1.18-10.52), which was associated with violence (OR = 3.68, 95% CI: 2.44-5.55) [31].

Additionally, in a study comprising 1410 patients with schizophrenia across 56 sites in the United States, the 6-month prevalence of any violence was 19.1%, with 3.6% of participants reporting serious violent behaviour. It was shown that positive psychotic symptoms, such as persecutory ideation increased the risk of violence, while negative symptoms such as social

withdrawal decreased the risk. Moreover, serious violence was associated with psychotic and depressive symptoms, as well as childhood conduct issues, and prior victimisation [32].

Similarly, in a multinational case-control study comparing patients with schizophrenia in forensic settings, comprising 221 individuals with a lifetime history of serious interpersonal violence, relative to 177 patients without a history of violence, forensic patients showed a greater prevalence of comorbid personality disorder (29.3% v. 7.6%), and were more likely to be exposed to severe violence during childhood [22]. Higher levels of disability, poorer performance in cognitive speed tasks, as well as lower social functioning were found to be protective factors, perhaps as a proxy measure of negative symptoms, alongside years of education [33].

Furthermore, as reported in a recent meta-analysis [34], large-scale studies using unaffected sibling controls have been conducted to more carefully adjust for confounding familial factors including genetic liabilities, and early environmental considerations. For instance, a study comprising 24,297 individuals with schizophrenia spectrum disorder [24], matched to sibling and general population controls, showed an increased odds of 1.8 (95% CI: 1.7-1.9) for violent crime in unaffected siblings, relative to general population controls, suggesting potential familial confounding factors [24]. These findings have also been related in sibling control studies conducted in Sweden [35], and Israel [36].

In terms of criminal recidivism in offenders with psychosis, a systematic review and meta-analysis comprising 3511 repeat offenders with psychotic disorders, 5446 individuals with other psychiatric disorders, and 71552 healthy individuals, showed a significantly increased risk of repeat offending in psychosis (pooled OR = 1.6, 95% CI = 1.4-18), although substantial heterogeneity was found, and this analysis was based on a subset of four studies [37]. A recent review also highlights that psychotic and manic symptoms are associated with an increased

likelihood of arrest for criminal offenses, although this appears to be driven by factors other than symptom severity [38].

Furthermore, with respect to criminal outcomes, available evidence suggests that one in eight men, and one in sixteen women will subsequently commit a serious criminal offense after release from a psychiatric facility [39]. This phenomenon is not isolated to specific geographical or generational effects, considering that in a systematic review comprising 33,588 individuals from 24 countries and 109 datasets, high rates of mental illness in prisoners were found in both high- and low-income countries over the timespan of four decades [40].

Additionally, results from a large Swedish registry study comprising 98,082 individuals with a history of hospitalization suggests that one in every twenty violent crimes is committed by someone with severe mental illness [41]. Given the high prevalence of criminal acts committed across cultures in individuals with severe mental illness, there has been a concerted effort to identify predictors of prospective criminal risk following discharge from psychiatric facilities.

1.2.5. Limitations of current methods of risk prediction

In response to this, actuarial assessments became increasingly widespread, which use statistical algorithms to identify prospective patient risk, usually at the group level [42]. However, there is little evidence that actuarial risk estimates can accurately determine whether a specific patient will reoffend or commit subsequent acts of violence [43].

Among the existing actuarial risk assessment methods, which assess the likelihood of violence across a group within a certain window of time [44], methods such as the Violent Risk Appraisal Guide (VRAG) have shown an area under the receiver operating characteristic curve (AUC) of 0.703 in identifying prospective criminal recidivism [45], with a slightly higher AUC when

examining patients who have committed prior violent offences (AUC=0.763). Additionally, the VRAG has shown a median AUC of 0.69 in predicting community violence in patients with schizophrenia across two studies [46]. Similarly, the Historical Clinical and Risk Management - 20 (HCR-20) is another structured tool to assess the risk of violence, that shows an AUC ranging from 0.674-0.723 in predicting prospective aggressive behaviour in men with schizophrenia living in the community [47].

However, available actuarial risk assessment in forensic settings carry several limitations, including assuming a linear additive relationship between variables in predicting a complex outcome such as physical aggression in schizophrenia. While it can be argued that additive approaches to risk prediction that assume equal weightings between risk factors are not necessarily a limitation, provided they show some utility in clinical contexts, it is important to provide further context. For instance, many of the predictors used in actuarial risk assessments show high correlations with each other, and as such, there is a potential of moderate to large degree of multicollinearity, which may violate the assumption of independence between variables within multiple linear regression models [48]. For instance, within a study validating the VRAG-Revised (VRAG-R) in a sample of 120 adult male offenders, small to moderate multicollinearity was observed across items, and only a subset of variables were identified as significant predictors of violent recidivism [49]. Moreover, there is little evidence to suggest that the HCR-20, which utilizes historical risk factors as a structured professional judgement tool, is effective in assessing and managing the risk of violence [50]. Altogether, while it remains inconclusive whether linear or nonlinear approaches are more appropriate to model violent behavior in schizophrenia, strategies comparing these methods are warranted.

To further complicate matters, most risk prediction methods have not reported performance indicators, such as AUC, and significant variation has been observed across studies and risk instruments in reporting practices [51]. Moreover, among studies that have reported performance metrics, most have relied on AUC to assess model performance, which ignores the goodness-of-fit and predicted probability values of the model in detecting true positives (sensitivity) and true negatives (specificity). As such, the model may show reasonable AUC, but fail to meaningfully predict physical aggression or criminal outcomes in the sample [52].

Apart from AUC, an important consideration in determining the clinical viability of prognostic models are the true positive predictive values (PPV) and negative predictive values (NPV), in this case corresponding to the instances of physical aggression and non-aggression that are correctly identified, respectively. PPV is calculated as the true positive rate, divided by the sum of the true and false positive rate, multiplied by 100. Conversely, NPV is calculated as the true negative rate, divided by the sum of the true negative and false negative rate, multiplied by 100 [53]. In other words, the PPV is the probability an individual with a positive result, in this case predicted to be physically aggressive, who will prospectively engage in physical aggression. Similarly, NPV is the probability an individual with a negative result, predicted to be non-aggressive, will not engage in aggression at follow-up.

It is necessary to caution while a prognostic model will ideally show a PPV and NPV approximating 100%, this rarely occurs in practical terms, and unlike with sensitivity and specificity metrics, PPV and NPV are impacted by the prevalence, or base rate, of the condition/disease in question. As such, an optimal threshold of PPV and sensitivity, or NPV and specificity, depends largely on how common the condition occurs, as well as the costs of false positives relative to false negatives. Even in cases where the PPV of a model is low, this can be

useful if the costs of intervention in those with false-positive results are low, relative to the benefits in intervening among those with a true-positive result [54]. In the specific example of physical aggression in patients with schizophrenia, if the NPV of the test is high, negative predictions are useful to reject the presence of physical aggression in the sample, however in cases where the PPV is low, positive predictions of physical aggression have an increased likelihood of being false positives. As such, models with a lower PPV and higher NPV would be more likely to incorrectly classify patients as physically aggressive, while more correctly identifying patients who were not physically aggressive. Conversely, if the PPV of the test is high and NPV is low, positive predictions can be helpful to identify the presence of physical aggression, however with a low NPV, negative predictions have a higher probability of being a false negative. As such, these models would be more likely to incorrectly classify patients as non-aggressive, while more correctly identifying physically aggressive patients.

Whether one scenario is preferrable largely depends on the specific context of how the model is implemented, and the ramifications in clinical care for patients. In cases where patients who are predicted to be positive cases are simply triaged as high risk and monitored more closely, models with higher NPV and lower PPV may have greater practical utility, as they are more likely to correctly identify non-aggressive patients, who can therefore be considered as low risk. While false positives will inevitably emerge in a high NPV and low PPV model, flagged patients can be more closely monitored, to decrease the likelihood of physical aggression occurring. Considering that available risk prediction methods largely focus on AUC as the exclusive performance metric, it remains difficult to assess their relative ability to detect true and false positives and negatives. Altogether, available tools to prospectively predict short-term physical aggression among patients in forensic settings remain limited, even though this remains a pressing need in

the clinical care of these individuals [26]. Moreover, there is a need for new tools to predict criminal recidivism among individuals[55].

1.2.6. Supervised machine learning: classification models

Broadly speaking, supervised machine learning is a subcategory of artificial intelligence where the model attempts to learn representations from labeled training data, or a set of features, to predict a given outcome [56]. In classification tasks, this outcome is categorical in nature, for instance, discriminating individuals with schizophrenia from controls. Conversely, in regression tasks, this outcome is continuous in nature, for instance, predicting symptom change scores in response to ketamine treatment in major depressive disorder [57]. In terms of classification-based models, or classifiers, common types of algorithms include logic (symbolic) methods, such as decision trees, which use conditional logic, with a series of nodes (features) and branches (outcomes), where all features are considered in the training data, and the decision tree attempts to find an optimal split of nodes with the lowest cost function [58]; whereas statistical learning algorithms such as Bayesian networks, explicitly calculate the probability of a training instance belonging to a specific class, and instance-based learning, such as *k*-nearest neighbours (kNN), do not generate a series of abstractions from the underlying training data, but instead generate classification predictions using the specific instances themselves [59].

Common algorithms used in classifiers include linear models, tree-based models, and kernel-based methods. Linear models, such as logistic regression [60], are a form of linear regression with a sigmoid function used to map probabilities between 0 and 1. Tree-based models include both bagging and boosting algorithms, which are a form of ensemble learners, that utilize decision trees [61]. Bagging involves a bootstrap procedure to generate multiple subsets of observations with replacements, where models are run independently and in parallel with each other, and final

predictions involve combining the predictions from all models, which decrease model variance relative to standard decision trees [62]. Conversely, boosting involves developing sequential weak hypotheses (learners) that involve simple decision trees with a few nodes, where in cases where an input is misclassified by a hypothesis, its weight is increased so that the subsequent hypothesis is more likely to correctly classify the instance. While initially all data points are given equal weight, a weighted average of iteratively fitted weak learners decreases model bias, and generally performs well in classification tasks [62]. Some examples of bagging algorithms include Random forest, bagged CART, and conditional forest, with random forest only selecting a subset of features at random out of the total [63], whereas bagged CART selects all features [62], and conditional forests use conditional inference trees as base learners, respectively [64] .

Furthermore, kernel-based algorithms, use a linear classifier to solve a non-linear problem, where input data is mapped into a higher dimensional space, to compute the dot product between our features and outcome, without explicitly computing this high-dimensional space [65]. Support Vector Machine (SVM) is the most used algorithm that incorporates a kernel function. In the absence of a kernel, SVM is a linear algorithm which includes a separating hyperplane that attempts to separate all samples with an optimal line with the greatest margin between classes [66]. However, since in most cases there will be several training instances on the incorrect side of the separating hyperplane, a soft margin is used to indicate the degree to which violations in the separating hyperplane are permissible. With the addition of a kernel function, data is projected from a low dimensional space to a higher dimensional space, to generate a more complex separating hyperplane with less violations, and ideally, greater model accuracy [66].

Furthermore, in terms of classification tasks, binary classification involves discriminating between two classes (e.g., Schizophrenia vs controls), whereas multiclass classification involves

discriminating between more than two classes (e.g., Schizophrenia vs Bipolar disorder vs controls), respectively [67]. Moreover, an important consideration in model development is hyperparameter tuning, which involves finding a configuration of tuning parameters prior to model training that results in the best performance (e.g., accuracy for classification models, and lowest root mean squared error for regression models, respectively). A more detailed overview of supervised machine learning [68], algorithm selection [69], and hyperparameter tuning [69] can be found elsewhere.

1.2.7. Ethical challenges of predictive models in forensic psychiatry patients

Machine learning models raise a variety of opportunities and avenues to develop educational tools, preventive measures, and shape public policy [70]. However, despite the potential for improving our ability to prognosticate clinical outcomes, in the context of forensic psychiatry, it is important to be cognizant of reducing the potential for further stigmatizing these vulnerable individuals, while also respecting their rights, as well as enhancing their safety and well-being.

Several pertinent questions arise when evaluating the utility and implementation of such algorithms. For instance, an important consideration that is often overlooked is model interpretability. So called "black box" methods may perform well in testing and validation datasets, however without a rudimentary understanding of the directionality, and interaction effects, of important features, we lack the transparency required to justify implementing these models in high stakes clinical settings [71]. Toward this end, new methods leveraging the internal structure of tree based algorithms can be used to directly measure local feature interaction effects, and provide insight into the magnitude, prevalence, and direction of a feature's effect [72].

Similarly, even among classification models that demonstrate high accuracy, there will be instances where individuals are misclassified. In cases where the risks of misclassification are

low, this may be largely unimportant. However, when dealing with the complex intersectionality between healthcare, personal freedom, and societal risk, this becomes a challenging consideration. For instance, how can we introduce ethical constraints in our models without significantly impacting their overall accuracy and utility? While this remains open to debate, it is important to ensure that our models are not predicated on immutable characteristics, and ensuring free, informed, and ongoing consent [73]. Moreover, meaningfully engage with stakeholders (healthcare providers, patients, and their families) will likely be required to reasonably implement predictive models into clinical care, to ensure the scope of the problem, and important ethical considerations, are adequately elucidated.

1.3. Predicting treatment response in psychiatric disorders

1.3.1. Treatment response prediction using EEG in major depressive disorder

Apart from prognostic machine learning models used to predict meaningful clinical outcomes in psychiatric disorders, such as prospective physical aggression in schizophrenia, and violent and criminal behaviors in individuals with psychiatric conditions more broadly, there has been an increasing focus in the field on predicting treatment response to medications and interventions at an individual level. In the context of MDD, it was notably demonstrated in the Sequential Treatment Alternatives to Relieve Depression (STAR*D) study that antidepressants fail to facilitate remission in most patients with major depressive disorder (MDD), and that there is no clearly preferred medication when patients inadequately respond to several courses of antidepressants [74]. Similarly, data from a multicentre randomised controlled trial spanning 2439 patients across 73 general practices in the United Kingdom found that 55% of patients (95% CI: 53-58%) met the threshold for treatment resistant depression, defined as ≥14 on the BDI-II, and who had been taking antidepressant medication of an adequate dose, for at least 6 weeks [75].

In contrast to neuroimaging modalities such as fMRI and MRI, which show a high cost associated with each scan, and excessive wait times to access a limited number of MRI machines, electroencephalography (EEG) is comparably more cost-effective and scalable as a potential clinical tool to predict treatment response. As described elsewhere [76], EEG oscillations refer to rhythmic electrical activity in the brain and constitute a mechanism where the brain can regulate changes within selected neuronal networks. This repetitive brain activity emerges because of the interactions of large populations of neurons. As such, there is evidence that MDD may be related to abnormalities in large-scale cortical and subcortical systems distributed across frontal, temporal, parietal, and occipital regions [76].

For instance, power amplitudes in specific frequency bands, known as band power, are associated with different mechanisms in the brain. Although incompletely understood, alpha band power (8-12 Hz) reflects sensory and attentional inhibition and has been shown to be associated with creative ideation [77], beta frequencies (13-30 Hz) are prominent during problem solving [78], while delta frequencies (≤4 Hz) are notable during deep sleep [79], gamma frequencies (30-80 Hz) during intensive concentration [80], and greater theta band frequencies (4-8 Hz) during relaxation, respectively [81]. Alpha asymmetry, which measures the relative alpha band power between hemispheres, particularly within frontal electrodes, have been shown to discriminate individuals with MDD from healthy controls, although inconsistencies have been found across literature [82]. Similarly, beta and low gamma powers in fronto-central regions have been shown to be negatively correlated with inattention scores in MDD [83]. Moreover, intrinsic local beta oscillations in the subgenual cingulate were found to be inversely related to depressive symptoms, particularly in the lower beta range of ~13-25 Hz [84]. Additionally, in specific contexts, gamma rhythms, which represent neural oscillations between 25 and 140 Hz, have been

shown to distinguish patients with MDD from healthy controls, and various therapeutic agents for depression have also been shown to alter gamma oscillations [85]. Patients with depression also show more random network structure, and differences in signal complexity [83], which may serve as replicable biomarkers of treatment response and remission.

1.3.2. Data-driven biomarkers of treatment response in randomized clinical trials

In addition to cost-effective measures such as EEG, there have been several predictive models thus far using baseline biological data within randomized clinical trials to predict treatment response. This strategy, unlike treatment response prediction in the absence of a comparator arm or placebo control, provides an opportunity to assess whether there are data-driven biomarkers specific to response to a given intervention. Considering that individual patients may deviate from the average group response, it can be expected that a specific treatment with demonstrated efficacy, relative to placebo, may not be efficacious across all patients. Additionally, due to strict inclusion/exclusion criteria meta-analyses and randomized controlled trials (RCTs) cannot properly map the complexity that are often seen in real patients, and as a result, are unable to render tailor-made evidence [86]. In fact, the very idiosyncrasies that characterise most patients, such as multimorbidity profiles, are often exclusion criteria in clinical trials.

It is also important to mention that statistically significant associations at the aggregate level do not necessarily translate into clinical benefit. For instance, in a network meta-analysis comparing the efficacy and acceptability of 21 antidepressant drugs across 522 trials for the acute treatment of adults with Major Depressive Disorder (MDD), while all antidepressants were found to be more efficacious than placebo, significant variability in efficacy and acceptability was observed between medications in head-to-head trials [87]. Similar heterogeneity in treatment efficacy was also observed across patients with schizophrenia in a network meta-analysis comprising 402

trials and 32 oral antipsychotics, with large differences in side effects between medications [88]. Altogether, available evidence suggests that approximately 20-60% of patients with psychiatric disorders continue to show significant residual symptoms following a course of treatment of sufficient dose and duration [89].

Despite clinical heterogeneity in response to medications that have been shown to be effective in randomized placebo-controlled trials, we currently lack objective biomarkers to guide the clinical likelihood of sufficient symptomatic improvement, inadequate symptom reduction, or remission within a specific patient to a given course of treatment. As such, patients continue to endure prolonged periods of "trial-and-error" in search of effective treatment and the burden associated with this process. Moreover, validated, and reliable biomarkers are needed to improve our understanding of the mechanisms of patient remission in response to specific treatments. For instance, while first-line antidepressants such as fluoxetine have been shown to be effective in many patients with depression for over 3 decades [90], debate remains surrounding their exact mechanisms of action [91]. Therefore, new strategies are required to determine which treatments are likely to be effective for a given patient, expedite biomarker discovery, and improve our mechanistic understanding of how currently approved medications improve symptoms, to guide the development of next-generation therapeutics in psychiatry.

Towards this end, machine learning, as described in section 1.14, is a subfield of artificial intelligence focused on computational methods that can extract relevant information from complex datasets [92]. Such methods can model patterns to generate individualized predictions using high quality data from various modalities, such as neuroimaging, genetics, neurophysiology, and clinical features [93]. Incorporating these techniques into less restricted clinical trials with medications that have already proven their efficacy in previous RCTs will aid

in the development of precision psychiatry, by enabling more precise interventions that include patient's idiosyncrasies [94]. Considering the limitations of a "trial-and-error" approach to treatment in psychiatry, there is a major unmet need for individualized predictions of response to treatment within randomized clinical trials.

1.4. Main Aims

Due to the clinical challenges of predicting criminal and violent outcomes in patients with psychotic disorders, and identifying whether a given individual will respond to a specific course of treatment, we sought to: 1) predict longitudinal physical aggression in patients with schizophrenia, 2) systematically review and meta-analyze machine learning models to predict criminal and violent outcomes in patients, 3) discuss the ethical considerations of such predictive models, 4) discuss the utility of an emerging fast-acting antidepressant in MDD and the need for candidate biomarkers, 5) systematically review and meta-analyze predicting treatment response using electroencephalography (EEG) in MDD, and 6) systematically review data-driven biomarkers of treatment response in randomized clinical trials in psychiatry.

1.5. Specific Objectives

The specific objectives of this thesis were to:

1) In chapter 2, predict longitudinal physical aggression in patients with Schizophrenia, within forensic settings, at an individual level using routinely collected clinical variables,

2) In chapter 3, provide a systematic review and meta-analysis of machine learning models to predict criminal and violent outcomes in psychiatry,

3) In chapter 4, discuss ethical considerations of developing predictive models of criminal and violent outcomes in psychiatry,

4) In chapter 5, discuss the potential of nasal esketamine as a fast-acting antidepressant, and identify the major unmet need for candidate biomarkers to predict treatment response to esketamine at an individual level,

5) In chapter 6, provide a systematic review and meta-analysis of predicting treatment response using EEG in major depressive disorder (MDD),

6) In chapter 7, provide a systematic review of data-driven biomarkers of treatment response within randomized clinical trials in psychiatry,

1.6. Hypotheses

The hypotheses for objectives 1, 2, and 6 are as follows:

1) Routinely collected baseline clinical variables will predict longitudinal physical aggression in patients with schizophrenia,

2) Machine learning models incorporating evidence-based risk factors can predict criminal and violent outcomes in individuals with psychiatric disorders,

3) Predictive models of treatment response using EEG will show better performance in neurostimulation trials, relative to pharmacological trials, across patients with MDD.

1.7. References

1. Salazar de Pablo, G. *et al.* Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophr. Bull.* **47**, 284–297 (2021).
2. Cuthbert, B. N. The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry* **13**, 28–35 (2014).
3. Insel, T. R. & Cuthbert, B. N. Brain disorders? Precisely: Precision medicine comes to psychiatry. *Science (80-.).* (2015) doi:10.1126/science.aab2358.
4. Charlson, F. J. *et al.* Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophr. Bull.* **44**, 1195–1203 (2018).
5. Saha, S., Chant, D., Welham, J. & McGrath, J. A Systematic Review of the Prevalence of Schizophrenia. *PLoS Med.* **2**, e141 (2005).
6. Fletcher, P. C. & Frith, C. D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* **10**, 48–58 (2009).
7. Correll, C. U. & Schooler, N. R. Negative Symptoms in Schizophrenia: A Review and Clinical Guide for Recognition, Assessment, and Treatment. *Neuropsychiatr. Dis. Treat.* **Volume 16**, 519–534 (2020).
8. Foussias, G. & Remington, G. Negative Symptoms in Schizophrenia: Avolition and Occam's Razor. *Schizophr. Bull.* **36**, 359–369 (2010).
9. Månsson, K. N. T. *et al.* Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. *Transl. Psychiatry* **5**, e530 (2015).
10. Wobrock, T. *et al.* Cognitive impairment of executive function as a core symptom of schizophrenia. *World J. Biol. Psychiatry* **10**, 442–451 (2009).
11. Keefe, R. S. E. & Fenton, W. S. How Should DSM-V Criteria for Schizophrenia Include Cognitive Impairment? *Schizophr. Bull.* **33**, 912–920 (2007).
12. Zipursky, R. B. Why Are the Outcomes in Patients With Schizophrenia So Poor? *J. Clin. Psychiatry* **75**, 20–24 (2014).
13. Harrow, M. Do Patients with Schizophrenia Ever Show Periods of Recovery? A 15-Year Multi-Follow-up Study. *Schizophr. Bull.* **31**, 723–734 (2005).
14. Penttilä, M., Jääskeläinen, E., Hirvonen, N., Isohanni, M. & Miettunen, J. Duration of untreated psychosis as predictor of long-term outcome in schizophrenia: Systematic review and meta-analysis. *Br. J. Psychiatry* **205**, 88–94 (2014).
15. Perkins, D. O., Gu, H., Boteva, K. & Lieberman, J. A. Relationship Between Duration of Untreated Psychosis and Outcome in First-Episode Schizophrenia: A Critical Review and Meta-Analysis. *Am. J. Psychiatry* **162**, 1785–1804 (2005).

16. Marshall, M. *et al.* Association Between Duration of Untreated Psychosis and Outcome in Cohorts of First-Episode Patients. *Arch. Gen. Psychiatry* **62**, 975 (2005).

17. Siegel, S. J. *et al.* Prognostic Variables at Intake and Long-Term Level of Function in Schizophrenia. *Am. J. Psychiatry* **163**, 433–441 (2006).

18. Barlati, S. *et al.* Primary and secondary negative symptoms severity and the use of psychiatric care resources in schizophrenia spectrum disorders: A 3-year follow-up longitudinal retrospective study. *Schizophr. Res.* **250**, 31–38 (2022).

19. Sevy, S., Nathanson, K., Visweswaraiah, H. & Amador, X. The relationship between insight and symptoms in schizophrenia. *Compr. Psychiatry* **45**, 16–19 (2004).

20. Lysaker, P., Bell, M., Milstein, R., Bryson, G. & Beam-Goulet, J. Insight and Psychosocial Treatment Compliance in Schizophrenia. *Psychiatry* **57**, 307–315 (1994).

21. O'Donnell, C. Compliance therapy: a randomised controlled trial in schizophrenia. *BMJ* **327**, 834–0 (2003).

22. Buckley, P. F. *et al.* Insight and Its Relationship to Violent Behavior in Patients With Schizophrenia. *Am. J. Psychiatry* **161**, 1712–1714 (2004).

23. Ekinci, O. & Ekinci, A. Association between insight, cognitive insight, positive symptoms and violence in patients with schizophrenia. *Nord. J. Psychiatry* **67**, 116–123 (2013).

24. Fazel, S., Wolf, A., Palm, C. & Lichtenstein, P. Violent crime, suicide, and premature mortality in patients with schizophrenia and related disorders: a 38-year total population study in Sweden. *The Lancet Psychiatry* **1**, 44–54 (2014).

25. Coid, J. W. *et al.* The Relationship Between Delusions and Violence. *JAMA Psychiatry* **70**, 465 (2013).

26. Faay, M. D. M. & Sommer, I. E. Risk and Prevention of Aggression in Patients With Psychotic Disorders. *Am. J. Psychiatry* **178**, 218–220 (2021).

27. Whiting, D., Gulati, G., Geddes, J. R. & Fazel, S. Association of Schizophrenia Spectrum Disorders and Violence Perpetration in Adults and Adolescents From 15 Countries. *JAMA Psychiatry* **79**, 120 (2022).

28. Buchanan, A. *et al.* Correlates of Future Violence in People Being Treated for Schizophrenia. *Am. J. Psychiatry* (2019) doi:10.1176/appi.ajp.2019.18080909.

29. Moulin, V. *et al.* Impulsivity in early psychosis: A complex link with violent behaviour and a target for intervention. *Eur. Psychiatry* **49**, 30–36 (2018).

30. Storvestre, G. B. *et al.* Childhood Trauma in Persons With Schizophrenia and a History of Interpersonal Violence. *Front. Psychiatry* **11**, (2020).

31. Keers, R., Ullrich, S., DeStavola, B. L. & Coid, J. W. Association of Violence With Emergence of Persecutory Delusions in Untreated Schizophrenia. *Am. J. Psychiatry* **171**, 332–339 (2014).

32. Swanson, J. W. *et al.* A National Study of Violent Behavior in Persons With

Schizophrenia. *Arch. Gen. Psychiatry* **63**, 490 (2006).

33. de Girolamo, G. *et al.* A multinational case−control study comparing forensic and non-forensic patients with schizophrenia spectrum disorders: the EU-VIORMED project. *Psychol. Med.* 1–11 (2021) doi:10.1017/S0033291721003433.

34. Whiting, D., Lichtenstein, P. & Fazel, S. Violence and mental disorders: a structured review of associations by individual diagnoses, risk factors, and risk assessment. *The Lancet Psychiatry* **8**, 150–161 (2021).

35. Sariaslan, A., Larsson, H. & Fazel, S. Genetic and environmental determinants of violence risk in psychotic disorders: a multivariate quantitative genetic study of 1.8 million Swedish twins and siblings. *Mol. Psychiatry* **21**, 1251–1256 (2016).

36. Fleischman, A., Werbeloff, N., Yoffe, R., Davidson, M. & Weiser, M. Schizophrenia and violent crime: a population-based study. *Psychol. Med.* **44**, 3051–3057 (2014).

37. Fazel, S. & Yu, R. Psychotic Disorders and Repeat Offending: Systematic Review and Meta-analysis. *Schizophr. Bull.* **37**, 800–810 (2011).

38. Lamberti, J. S., Katsetos, V., Jacobowitz, D. B. & Weisman, R. L. Psychosis, Mania and Criminal Recidivism: Associations and Implications for Prevention. *Harv. Rev. Psychiatry* **28**, 179–202 (2020).

39. Arboleda-Flórez, J. Forensic psychiatry: contemporary scope, challenges and controversies. *World Psychiatry* (2006).

40. Coid, J., Mickey, N., Kahtan, N., Zhang, T. & Yang, M. Patients discharged from medium secure forensic psychiatry services: Reconvictions and risk factors. *Br. J. Psychiatry* (2007) doi:10.1192/bjp.bp.105.018788.

41. Fazel, S. & Seewald, K. Severe mental illness in 33 588 prisoners worldwide: Systematic review and meta-regression analysis. *British Journal of Psychiatry* (2012) doi:10.1192/bjp.bp.111.096370.

42. Litwack, T. R. Actuarial versus clinical assessments of dangerousness. *Psychol. Public Policy, Law* (2001) doi:10.1037/1076-8971.7.2.409.

43. Hart, S. D., Michie, C. & Cooke, D. J. Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *Br. J. Psychiatry* (2007) doi:10.1192/bjp.190.5.s60.

44. Singh, J. P. & Fazel, S. Forensic Risk Assessment. *Crim. Justice Behav.* **37**, 965–988 (2010).

45. Kröner, C., Stadtland, C., Eidt, M. & Nedopil, N. The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. *Crim. Behav. Ment. Heal.* **17**, 89–100 (2007).

46. Singh, J. P., Serper, M., Reinharth, J. & Fazel, S. Structured Assessment of Violence Risk in Schizophrenia and Other Psychiatric Disorders: A Systematic Review of the Validity, Reliability, and Item Content of 10 Available Instruments. *Schizophr. Bull.* **37**, 899–912

(2011).

47. Michel, S. F. *et al.* Using the HCR-20 to Predict Aggressive Behavior among Men with Schizophrenia Living in the Community: Accuracy of Prediction, General and Forensic Settings, and Dynamic Risk Factors. *Int. J. Forensic Ment. Health* **12**, 1–13 (2013).

48. MAXWELL, A. E. Limitations on the use of the multiple linear regression model. *Br. J. Math. Stat. Psychol.* **28**, 51–62 (1975).

49. Glover, A. J. J., Churcher, F. P., Gray, A. L., Mills, J. F. & Nicholson, D. E. A cross-validation of the Violence Risk Appraisal Guide—Revised (VRAG–R) within a correctional sample. *Law Hum. Behav.* **41**, 507–518 (2017).

50. Challinor, A., Ogundalu, A., McIntyre, J. C., Bramwell, V. & Nathan, R. The empirical evidence base for the use of the HCR-20: A narrative review of study designs and transferability of results to clinical practice. *Int. J. Law Psychiatry* **78**, 101729 (2021).

51. Singh, J. P., Desmarais, S. L. & Van Dorn, R. A. Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review. *Behav. Sci. Law* **31**, 55–73 (2013).

52. Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2008).

53. Skaik, Y. Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* **56**, 341 (2008).

54. Kessler, R. C. Clinical Epidemiological Research on Suicide-Related Behaviors—Where We Are and Where We Need to Go. *JAMA Psychiatry* **76**, 777 (2019).

55. Lin, Z. "Jerry", Jung, J., Goel, S. & Skeem, J. The limits of human predictions of recidivism. *Sci. Adv.* **6**, (2020).

56. Gianey, H. K. & Choudhary, R. Comprehensive Review On Supervised Machine Learning Algorithms. in *Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017* vols 2018-January 38–43 (Institute of Electrical and Electronics Engineers Inc., 2018).

57. Kotsiantis, S. B., Zaharakis, I. D. & Pintelas, P. E. Machine learning: A review of classification and combining techniques. *Artif. Intell. Rev.* (2006) doi:10.1007/s10462-007-9052-3.

58. Kingsford, C. & Salzberg, S. L. What are decision trees? *Nat. Biotechnol.* **26**, 1011–1013 (2008).

59. Aha, D. W., Kibler, D. & Albert, M. K. Instance-based learning algorithms. *Mach. Learn.* **6**, 37–66 (1991).

60. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).

61. González, S., García, S., Del Ser, J., Rokach, L. & Herrera, F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools,

performance study, practical perspectives and opportunities. *Inf. Fusion* **64**, 205–237 (2020).

62. Sutton, C. D. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics* vol. 24 303–329 (2005).

63. Breiman, L. Random forests. *Mach. Learn.* (2001) doi:10.1023/A:1010933404324.

64. Nasejje, J. B., Mwambi, H., Dheda, K. & Lesosky, M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med. Res. Methodol.* **17**, 115 (2017).

65. Hofmann, T., Schölkopf, B. & Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **36**, (2008).

66. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* (2006) doi:10.1038/nbt1206-1565.

67. Grandini, M., Bagli, E. & Visani, G. Metrics For Multi-Class Classification: An Overview. *arXiv* (2020).

68. Osarogiagbon, A. U., Khan, F., Venkatesan, R. & Gillard, P. Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. *Process Saf. Environ. Prot.* **147**, 367–384 (2021).

69. Yu, T. & Zhu, H. Hyper-Parameter Optimization: A Review of Algorithms and Applications. (2020).

70. Passos, I. C., Mwangi, B. & Kapczinski, F. Big data analytics and machine learning: 2015 and beyond. *The Lancet Psychiatry* (2016) doi:10.1016/S2215-0366(15)00549-0.

71. Rudin, C. models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, (2019).

72. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* (2020) doi:10.1038/s42256-019-0138-9.

73. Nozick, R. *Philosophical Explanations*. vol. 92 (Harvard University Press, 1981).

74. Trivedi, M. H. *et al.* Evaluation of Outcomes With Citalopram for Depression Using Measurement-Based Care in STAR * D : Implications for Clinical Practice. *Am. J. Psychiatry* 28–40 (2006).

75. Thomas, L. *et al.* Prevalence of treatment-resistant depression in primary care: cross-sectional data. *Br. J. Gen. Pract.* **63**, e852–e858 (2013).

76. Fingelkurts, A. A. & Fingelkurts, A. A. Altered Structure of Dynamic Electroencephalogram Oscillatory Pattern in Major Depression. *Biol. Psychiatry* **77**, 1050–1060 (2015).

77. Fink, A. & Benedek, M. EEG alpha power and creative ideation. *Neurosci. Biobehav. Rev.* **44**, 111–123 (2014).

78. Roslan, N. S., Amin, H. U., Izhar, L. I., Saad, M. N. M. & Sivapalan, S. Role of EEG delta and beta oscillations during problem solving tasks. in *2016 6th International Conference on Intelligent and Advanced Systems (ICIAS)* 1–4 (IEEE, 2016). doi:10.1109/ICIAS.2016.7824138.

79. Amzica, F. & Steriade, M. Electrophysiological correlates of sleep delta waves. *Electroencephalogr. Clin. Neurophysiol.* **107**, 69–83 (1998).

80. Lally, N. *et al.* Glutamatergic correlates of gamma-band oscillatory activity during cognition: A concurrent ER-MRS and EEG study. *Neuroimage* **85**, 823–833 (2014).

81. Jacobs, G. D. & Friedman, R. EEG Spectral Analysis of Relaxation Techniques. *Appl. Psychophysiol. Biofeedback* **29**, 245–254 (2004).

82. Soares, F., Neto, D. A., Luís, J. & Rosa, G. Depression biomarkers using non-invasive EEG : A review. *Neurosci. Biobehav. Rev.* **105**, 83–93 (2019).

83. Roh, S.-C., Park, E.-J., Shim, M. & Lee, S.-H. EEG beta and low gamma power correlates with inattention in patients with major depressive disorder. *J. Affect. Disord.* **204**, 124–130 (2016).

84. Clark, D. L., Brown, E. C., Ramasubbu, R. & Kiss, Z. H. T. Intrinsic Local Beta Oscillations in the Subgenual Cingulate Relate to Depressive Symptoms in Treatment-Resistant Depression. *Biol. Psychiatry* **80**, e93–e94 (2016).

85. Fitzgerald, P. J. & Watson, B. O. Gamma oscillations as a biomarker for major depression: an emerging topic. *Transl. Psychiatry* **8**, 177 (2018).

86. Beckmann, J. S. & Lew, D. Reconciling evidence-based medicine and precision medicine in the era of big data: Challenges and opportunities. *Genome Med.* (2016) doi:10.1186/s13073-016-0388-7.

87. Cipriani, A. *et al.* Articles Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder : a systematic review and network meta-analysis. *Lancet* **391**, 1357–1366 (2018).

88. Leucht, S. *et al.* Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: A multiple-treatments meta-analysis. *Lancet* **382**, 951–962 (2013).

89. Howes, O. D., Thase, M. E. & Pillinger, T. Treatment resistance in psychiatry: state of the art and new directions. *Molecular Psychiatry* (2021) doi:10.1038/s41380-021-01200-3.

90. López-Muñoz, F. & Alamo, C. *Monoaminergic Neurotransmission: The History of the Discovery of Antidepressants from 1950s Until Today*. *Current Pharmaceutical Design* vol. 15 (2009).

91. Harmer, C. J., Duman, R. S. & Cowen, P. J. How do antidepressants work? New perspectives for refining future treatment approaches. *The Lancet Psychiatry* vol. 4 409–418 (2017).

92. Fan, J., Han, F. & Liu, H. Challenges of Big Data analysis. *National Science Review* (2014) doi:10.1093/nsr/nwt032.

93. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* (2015) doi:10.1038/nature14541.

94. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* (2014) doi:10.1186/2047-2501-2-3.

Chapter 2 - The HARM models: Predicting longitudinal physical aggression in patients with schizophrenia at an individual level

Authors: Devon Watts, MSc[1]; Mini Mamak, PhD[2]; Heather Moulden, PhD[2]; Casey Upfold[2]; Taiane de Azevedo Cardoso, MSc, PhD[2]; Flavio Kapczinski MSc, MD, PhD, FRCPC[1,2,3]; Gary Chaimowitz MB, ChB, MBA, FRCPC[2]

1. Neuroscience Graduate Program, McMaster University, Hamilton, Canada

2. Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada.

3. Instituto Nacional de Ciência e Tecnologia Translacional em Medicina (INCT-TM), Porto Alegre, Brazil

*Corresponding author:

Flavio Kapczinski, MSc, MD, PhD, FRCP(C)

Professor, Psychiatry & Behavioural Neurosciences, McMaster University, 100 West 5th Street, Hamilton, ON L9C 0E3, 905.522.1155 Ext.35420, Email: flavio.kapczinski@gmail.com

Email for all other authors:

Devon Watts: wattsd21@gmail.com

Mini Mamak: mamakm@stjosham.on.ca

Heather Moulden: hmoulden@stjosham.on.ca

Casey Upfold: cupfold@stjosham.on.ca

Taiane de Azevedo Cardoso: taianeacardoso@gmail.com

Gary Chaimowitz: chaimow@mcmaster.ca

This chapter has been accepted in the **Journal of Psychiatric Research**

ABSTRACT

The prediction and prevention of aggression in individuals with schizophrenia remains a top priority within forensic psychiatric settings. While risk assessment methods are well rooted in forensic psychiatry, there are no available tools to predict longitudinal physical aggression in patients with schizophrenia within forensic settings at an individual level. In the present study, we used evidence-based risk and protective factors assessed at baseline, to predict prospective incidents of physical aggression (4-month, 12-month, and 18-month follow-up) among 151 patients with schizophrenia within the forensic mental healthcare system. Across our HARM models, the balanced accuracy (sensitivity + specificity / 2) of predicting physical aggressive incidents in patients with schizophrenia ranged from 59.73-87.33% at 4-month follow-up, 68.31-80.10% at 12-month follow-up, and 46.22-81.63% at 18-month follow-up, respectively. Additionally, we developed separate models, using clinician rated clinical judgement of short term and immediate violent risk, as a measure of comparison.

Several evidence-based modifiable predictors of prospective physical aggression in psychotic patients at an individual level, including changes in impulse control, substance abuse, impulsivity, treatment non-adherence, mood symptoms, substance abuse, psychotic symptoms, and poor family support. To the best of our knowledge, our HARM models are the first to predict longitudinal physical aggression at an individual level in patients with schizophrenia in forensic settings. However, it is important to caution that since these machine learning models were developed in the context of forensic settings, they may not be generalisable to individuals with schizophrenia more broadly. Moreover, considering the low base rate of physical aggressive incidents in the testing set (6.0-11.6% across timepoints), future studies with larger cohorts will be required to determine the replicability of these findings.

Keywords: machine learning; schizophrenia; artificial intelligence; psychotic disorders; precision psychiatry; computational neuroscience; criminality

INTRODUCTION:

The prediction and prevention of aggression in patients with psychotic disorders remains among the top priorities in their clinical care [1]. It has been shown that schizophrenia and related disorders are associated with substantially increased rates of violent crime [2]. In a recent meta-analysis and systematic review of the association between schizophrenia spectrum disorders and the perpetration of violence comprising 51,309 individuals across 24 studies in 15 countries over four decades [3], those with psychosis and comorbid substance misuse showed approximately 10-fold increased odds of prospective violence relative to general population controls. However, this relative risk was much lower, approximately 3-fold, among individuals lacking comorbidities [3]. Similarly, another meta-analysis comprising 204 studies across 166 independent datasets suggests that psychosis is associated with a 49-68% increased likelihood of violence, although substantial variability was found due to moderating factors such as how psychosis is measured, the presence of a comparison group, and study design [4]. As such, it is important to caution that there is little evidence to characterise individuals with schizophrenia as inherently dangerous [5], but rather, this highlights the necessity of identifying risk factors and applying preventative strategies among a subset of patients who show an elevated likelihood of prospective physical aggression.

Several prior studies have examined modifiable and causal risk factors including treatment nonadherence [6], impulsivity [7], and childhood trauma [8]. For instance, in a prospective longitudinal UK Prisoner Cohort Study [9], comprising individuals without psychosis (N=742), with schizophrenia (N=94), delusional disorder (N=29), and drug-induced psychosis (N=102), schizophrenia was found to be associated with violence, but only in the absence of treatment (Odds Ratio, OR = 3.76, 95% CI: 1.39-10.19). Moreover, untreated schizophrenia was associated with the appearance of persecutory delusions at follow-up (OR = 3.52, 95% CI: 1.18-10.52), which was associated with violence (OR = 3.68, 95% CI: 2.44-5.55) [9].

Additionally, in a study comprising 1410 patients with schizophrenia across 56 sites in the United States, the 6-month prevalence of any violence was 19.1%, with 3.6% of participants reporting serious violent behaviour. It was shown that positive psychotic symptoms, such as persecutory ideation increased the risk of violence, while negative symptoms such as social withdrawal decreased the risk. Moreover, serious violence was associated with psychotic and depressive symptoms, as well as childhood conduct issues, and prior victimisation [10].

Similarly, in a multinational case-control study comparing patients with schizophrenia in forensic settings, comprising 221 individuals with a lifetime history of serious interpersonal violence, relative to 177 patients without a history of violence, forensic patients showed a greater prevalence of comorbid personality disorder (29.3% v. 7.6%), and were more likely to be exposed to severe violence during childhood [11]. Higher levels of disability, poorer performance in cognitive speed tasks, as well as lower social functioning were found to be protective factors, perhaps as a proxy measure of negative symptoms, alongside years of education [11].

Furthermore, as reported in a recent meta-analysis [12], large-scale studies using unaffected sibling controls have been conducted to more carefully adjust for confounding familial factors including

genetic liabilities, and early environmental considerations. For instance, a study comprising 24,297 individuals with schizophrenia spectrum disorder [13], matched to sibling and general population controls, showed an increased odds of 1.8 (95% CI: 1.7-1.9) for violent crime in unaffected siblings, relative to general population controls, suggesting potential familial confounding factors [13]. These findings have also been related in sibling control studies conducted in Sweden [14], and Israel [15].

Despite the increased odds of physical violence and aggression among a subset of patients with schizophrenia, it remains a significant clinical challenge to predict which specific patients are likely to engage in violent acts before they occur. Considering that most patients with this condition do not show a lifetime history of violent offending and the low base rate of these events among those who do, patients with schizophrenia in forensic settings represent a high-risk group for prospective physical aggression.

Among the existing actuarial risk assessment methods, which assess the likelihood of violence across a group within a certain window of time [16], methods such as the Violent Risk Appraisal Guide (VRAG) have shown an area under the receiver operating characteristic curve (AUC) of 0.703 in identifying prospective criminal recidivism [17], with a slightly higher AUC when examining patients who have committed prior violent offences (AUC=0.763). Additionally, the VRAG has shown a median AUC of 0.69 in predicting community violence in patients with schizophrenia across two studies [18]. Similarly, the Historical Clinical and Risk Management - 20 (HCR-20) is another structured tool to assess the risk of violence, that shows an AUC ranging from 0.674-0.723 in predicting prospective aggressive behaviour in men with schizophrenia living in the community [19].

However, available actuarial risk assessment in forensic settings carry a number of limitations, including assuming a linear additive relationship between variables in predicting a complex outcome such as physical aggression in schizophrenia, and prior methods have relied on AUC to assess model performance, which ignores the goodness-of-fit and predicted probability values of the model in detecting true positives (sensitivity) and true negatives (specificity). As such, the model may show reasonable AUC, but fail to meaningfully predict physical aggression in the sample [20]. As such, available tools to prospectively predict short-term physical aggression among patients in forensic settings remain limited, even though this remains a pressing need in the clinical care of these individuals [1].

Increasingly, machine learning techniques have been used to make individualised predictions in various fields of healthcare [21]. In general, these algorithms can leverage existing datasets to detect patterns, and use these patterns to make predictions in independent datasets. These methods, alongside high-quality data, can be used to facilitate advancements in the diagnosis, assessment, and treatment of patients in psychiatry [22]. As such, in the present study, we developed a series of HARM models, using machine learning techniques alongside evidence-based static and modifiable risk factors, to predict longitudinal physical aggression (4-month, 12-month, and 18-month follow-up) in 151 at-risk patients with schizophrenia currently undergoing treatment within a forensic mental system.

2. Methods

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All analyses involving human patients were

approved by the Hamilton Integrated Research Ethics Board (#12857). Patient data were anonymized with digital identifiers removed, in line with ethical standards.

2.1. Study population

The study population comprised 151 patients diagnosed with Schizophrenia, according to the DSM-5 [23], undergoing treatment within a forensic psychiatry program in Canada. In Canada, individuals come under the jurisdiction of the forensic psychiatry system when they commit a criminal offence and are subsequently found not criminally responsible (NCR) or unfit to stand trial (UST) due to a mental disorder. All patients in the study were either NCR or UST at the time of data collection.

2.2. Measures

Data from the present study were gathered from the electronic Hamilton Anatomy of Risk Assessment (e-HARM), a structured professional judgement tool, developed for use in inpatient and outpatient psychiatric settings [24]. The e-HARM captures historical risk factors, including prior violent and nonviolent offences, major mental disorders, personality disorders, substance use, and cognitive deficits, alongside dynamic risk factors including rule adherence, patient insight, mood and psychotic symptoms, impulse control, social support, substance abuse, medication nonadherence, antisocial attitude, and stress. The e-HARM allows clinicians/clinical teams to easily consider empirically supported risk factors of violence, demographic information, protective factors, medications, psychiatric diagnoses, and then formulate risk estimates and risk management plans. Moreover, embedded in the e-HARM is the Aggressive Incidents Scale (AIS) [24], which provides a standardised method of recording aggressive incidents along a 9-point

scale in ascending order of severity, as shown in Supplementary Figure S1. Levels one through three comprise verbal aggression while levels four and higher involve physical aggression.

As discussed, elsewhere [25], the AIS strongly correlates with scores on the Modified Overt Aggression Scale (MOAS) (r= .92, p < .01), with considerable agreement between the AIS and MOAS (κ= .79, p < .0001) when dichotomizing aggression as present or absent. A list of all variables collected within the e-HARM, as well as candidate features considered in model development, can be found in Supplementary Table S3. Binary classification was used to dichotomize physically aggressive (AIS $\geq$ 4) and non-physically aggressive incidents (AIS <3) at follow-up timepoints.

Sixty-two variables, as detailed in Supplementary Table S3, were considered as candidate features within supervised binary machine learning classification models. This included all clinical, risk factors, protective factors, and treatment variables recorded within the eHARM, apart from variables related to clinician appraisal of immediate and short-term likelihood of patient violence, which were excluded from the candidate set of features. Considering that most of these variables were categorical, one-hot encoding was used to binarize factor levels. Nine machine learning models were compared, to dichotomize physical aggression and non-physical aggression at follow-up timepoints.

Each model was trained using baseline HARMs, corresponding to five assessments (Median = 88.50 days, SD = 32.12), to predict physical aggression at 4 months (Median = 114.50 days, SD = 41.79), 12 months (Median = 350 days, SD = 107.80), and 18 months (Median = 563.50 days, SD = 203.04), follow-up respectively. Further details can be found in the supplementary material. Importantly, our models were trained only using features available at baseline in the training set (60%), to predict longitudinal physical aggression at follow-up within a holdout

dataset (40%). Within the binary classification models, physical aggression was considered as the positive class, and non-aggression was considered as the negative class, respectively.

2.3. Machine Learning Algorithms

Nine machine learning algorithms (Boosted Logistic Regression, Elastic Net, Lasso Regression, k-nearest neighbours, Adaptive Boosting, Extreme Gradient Boosting, Random Forest, Bagged CART, and Conditional Forest) were implemented in R using various packages [25–28]. Features were centred and scaled using preProcess in Caret [26]. Zero and near-zero variance predictors were removed using the nearZeroVar function available in Caret [26]. Importantly, each of these algorithms incorporate slightly different regularisation parameters to address the issue of multicollinearity. One-hot encoding was used to transform categorical variables into dichotomous numerical values.

Briefly, boosted logistic regression involves adding a boosting parameter, or an ensemble of weak learners, to a linear model with a sigmoid function to reduce model bias [29]. Elastic net is a penalised least squares regression method that combines L1 and L2 regularisation from lasso and ridge methods [30]. This algorithm is efficient computationally and works well with highly correlated predictors. K-nearest neighbours is a simple and fast algorithm for classification, that involves tuning the number of nearest neighbours, with more defined boundaries as k is increased. However, since it is a non-parametric algorithm, meaning it does not make assumptions of the underlying data distribution, it scales poorly to larger datasets [31]. Furthermore, Adaptive boosting (AdaBoost) [32] and extreme gradient boosting (XGBoost) [33] both use a series of weak sequential learners, that are only slightly correlated with the true classification, and sequentially places greater weights on instances with incorrect predictions and high errors. However, XGBoost incorporates a specific implementation of gradient boosting that

uses both L1 and L2 regularisation to improve model generalisation, whereas AdaBoost involves an exponential loss function. Moreover, XGBoost is more computationally efficient, able to handle sparse data, and provides many hyperparameters that can be tuned to increase model performance [33]. Random forest, bagged CART, and conditional forest are all tree-based models, however, the fundamental difference is that random forest only selects a subset of features at random out of the total [34], whereas bagged CART selects all features [35], and conditional forests use conditional inference trees as base learners [36].

2.4. Feature selection

Within machine learning models, feature selection is an important pre-processing method to decrease dimensionality by removing irrelevant features [37,38]. A detailed overview of available feature selection methods can be found elsewhere [39]. In general, machine learning models tend to show greater generalizability in independent datasets when the number of features is limited. In the present study, most candidate features were categorical in nature. Considering this, embedded feature selection was used, which combines filter and wrapper methods, and selects a subset of features from the overall model that show the highest variable importance using the VarImp function in R. For instance, as discussed in section 2.6, in an elastic net model, variable importance is calculated using the absolute value of the coefficients within the tuned model. Moreover, in a random forest model, variable importance is calculated by computing the out-of-bag error rate for each tree, permuting each predictor variable, calculating the difference between these measures across all trees, and normalising by the standard deviation of the differences. In the model with the highest balanced accuracy for each outcome (4 month, 12-month, 18-month

follow-up), the top 30 features according to variable importance were retained, and performance was compared against an overall model comprising 67 variables. In all models, feature selection was only performed on training data (60%). Importantly, only variables available at baseline were considered as potential predictors. Variables containing 15% or more missing data were excluded. Furthermore, mean/median imputation was performed for numerical variables, and mode imputation was performed for categorical variables, respectively.

2.5. Addressing Class Imbalance

In binary classification problems, class imbalance is present when the number of a minority class (e.g., aggressive incidents) occurs much less often than in the majority class (e.g., non-aggressive incidents). Within the present study, as shown in Table 1, 9.5% of patients in the testing set committed an act of physical aggression (AIS score ≥ 4) at 4-month follow-up. Considering this, class imbalance was addressed by downsampling the majority class [40].

2.6. Model testing and validation

The HARM dataset was divided into training (60%) and testing sets (40%), respectively. This corresponded to 92 patients across baseline assessments (370 instances) in the training set, and 61 patients across follow-up in the testing set (181 instances), respectively. This training and testing threshold was selected considering the sparsity of aggressive patients at follow-up. As such, patients could commit more than one aggressive incidence during baseline or follow-up, with each instance recorded as a separate event. Within the testing set, there were 21 instances of physical aggression at 4-month follow-up (Median days since baseline = 114.50, SD = 41.79), 15

instances of physical aggression at 12-month follow-up (Median days = 350.00, SD = 107.80), and 11 instances at 18-month follow-up (Median days = 563.50, SD = 203.04), respectively.

Leave-one-group-out cross validation was used to estimate prediction error in the training set, which involves leaving one observation from each group out from the training set and predicting the response variable (physically aggressive vs. nonaggressive) in the left-out observations and calculating the mean standard error. Further details regarding the strengths and limitations of various cross-validation methods can be found elsewhere [41]. Model performance was assessed using the confusionMatrix function in R [27]. A confusion matrix is a table layout that provides an overview of model accuracy, misclassification rate, sensitivity, specificity, as well as true and false predictive values. This includes the number of correct and incorrect predictions, which are summarised with count values and broken down by each class. Further details can be found elsewhere [42].

To further evaluate the performance of the HARM models, and potential clinical utility, we developed separate models using clinician-rated estimates of the immediate and short-term likelihood of violence as input features. This was assessed using four variables, which evaluated the clinical likelihood of violence both immediately (days) and short-term (weeks) along a five-point scale, with 1 indicating low risk, and 5 indicating high risk, respectively. We also developed additional models, comprising both data-driven and clinical-likelihood of violence variables, as a comparator. Furthermore, to evaluate whether there were statistically significant differences in classifier performances, a McNemar's test with continuity correction was performed, between HARM vs clinical likelihood of violence (CLV) models, CLV vs combined models, and HARM vs combined models, respectively.

2.7. Model Interpretability

Variable importance plots were generated using the varImp function in the caret package in R [27]. Within boosted logistic regression and regularised logistic regression models, variable importance was calculated using the absolute value of the t-statistic for each model parameter. Similarly, in an elastic net model, this involved calculating the absolute value of the coefficients within the tuned model. Variable importance in the random forest model was calculated by computing the out-of-bag error rate for each tree, permuting each predictor variable, calculating the difference between these measures across all trees, and normalising by the standard deviation of the differences. Bagged and boosted tree models involve applying the same methodology as a single tree to all bootstrapped trees, and calculating total importance, and aggregating the importance over each boosting iteration, respectively. The kNN algorithm does not provide a method to calculate feature importance, and as such, this was not reported.

Results

In the present study, several machine learning models were developed, using features collected at baseline, to predict longitudinal physical aggression (4 months, 12 months, 18 months) in 153 patients with schizophrenia, in forensic settings, at an individual level. A summary of patient demographics can be found in Table 1. Across all algorithms, the balanced accuracy of predicting physical aggressive incidents in patients with schizophrenia at four-month follow-up ranged from 64.19-86.60%. The highest performance was observed within a random forest model (Balanced Accuracy = 86.60%; Accuracy = 87.33%, 95% CI: 82.21-91.41; PPV = 41.86; NPV = 98.31; AUC = 0.914 (95% CI: 0.872-0.951). Further information can be found in Table

2, and Supplementary Figure S2. Moreover, model sensitivity, corresponding to correctly predicting physical aggression in patients (true positive), was 85.71%, and model specificity, corresponding to correctly predicting non-aggression in patients (true negative), was 87.50%, respectively. Similar performance was also observed within a conditional forest model (Balanced Accuracy = 85.36%; Accuracy = 77.38%, 95% CI: 71.28-82.72), and elastic net model (Balanced Accuracy = 82.22%; Accuracy = 83.26%. Conversely, poorer performance was observed using kNN (Balanced Accuracy = 64.19%; Accuracy = 73.76%, 95% CI: 67.43-79.43), and lasso regression (Balanced Accuracy = 62.83%, Accuracy = 59.73% (95% CI: 52.94-66.25). However, it should be noted the accuracy of most models fell within a similar range of confidence intervals, which precludes a definitive statement as to the best performance.

Important features in the random forest model included worsening negative peer influence, worsening rule adherence, poor program participation, poor attitude, worsening stress management, substance abuse, and short-term escape risk, as well as currently being monitored for these behaviours. Similarly, changes in stress management, impulse control, worsening mood symptoms, worsening medication adherence, and frequency of medication use, were among the most important global features in the model. Further details regarding candidate features can be found in Supplementary Table S3.

Moreover, as shown in Supplementary Table S1, the balanced accuracy of predicting physical aggression in patients with schizophrenia in forensic settings at 12-month follow-up ranged from 65.58-86.15% across models. The best performance was observed within a random forest model, with a balanced accuracy of 86.15%, and overall accuracy of 80.10% (95% CI: 73.73-85.52). Additionally, the sensitivity of the model was 93.33%, and specificity was 78.97%, respectively. Important features at baseline in predicting 12-month physical aggression in the random forest

model, as shown in Supplementary Figure S3, included worsening impulse control, changes in rule adherence, worsening mood symptoms, worsening attitude, short-term escape risk, worsening program participation, poor stress management, use of haloperidol, presence of a personality disorder, and engagement in recreational and psychoeducational programs.

Additionally, as presented in Supplementary Table S2, the balanced accuracy of predicting physical aggression at 18-month follow-up was slightly lower, with a range of 46.22-81.81%. An XGBoost model showed the highest performance, with a balanced accuracy of 81.81%, overall accuracy of 83.43% (95% CI: 77.19-88.53), sensitivity of 80.00%, specificity of 83.62%, and AUC of 0.870 (95% CI: 0.814-0.918). Important features included worsening impulse control, changes in rule adherence over time, being highly engaged in a program/intervention, psychotic symptoms, poor stress management, changes in family support, and worsening substance abuse. Additional details can be found in Supplementary Figure S4.

Furthermore, across 4 month, 12 month, and 18 month follow-up timepoints, important baseline features predictive of subsequent aggression across models included change in attitude/cooperation (monitor and needs improvement), change in rule adherence (monitor and needs improvement), change in impulse control (monitor and needs improvement), change in stress management (monitor and needs improvement), worsening mood and psychotic symptoms, change in family support, presence of personality disorders, and peer influence.

Apart from the data-driven HARM models, we also sought to compare their performance relative to models using clinician-rated clinical likelihood of short-term and immediate violence. Moreover, we also assessed whether combining both data-driven HARM and clinician-rated likelihood of violence features lead to greater performance. As shown in Supplementary Table S4, a random forest model using clinician rated CLV showed the best performance, with a

sensitivity of 96.50%, specificity of 57.14%, PPV of 95.54, and NPV of 63.16, respectively. Additionally, as detailed in Supplementary Table S5, a random forest model combining both HARM and clinician rated CLV showed the best performance, with a sensitivity of 95.23%, specificity of 88.00%, PPV of 45.45, and NPV of 99.43, respectively.

Furthermore, a McNemar's test with continuity correction was used to assess whether statistically significant differences in error rates were observed across HARM, clinical judgement, and combined models, in respective pairwise combinations. No significant differences in error rates were observed between HARM and clinician rated CLV models (McNemar's chi-square (χ^2) = 2.37, p=0.123), or between HARM and combined models (McNemar's χ^2= 0.5, p=0.479). However, a significant difference was observed between combined and CLV models (McNemar's χ^2= 10.22, p= 0.001).

Discussion

To the best of our knowledge, this is the first study to longitudinally predict short-term physical aggression in patients with schizophrenia, in forensic settings, at an individual level. Importantly, the predictive models were developed using empirically supported risk factors of violence as candidate features, in conjunction with demographic variables, protective factors, and variables related to the course of treatment. Moreover, several potential protective factors emerged, including engagement in treatment programs, positive attitude, social support, family support, and medication adherence. Furthermore, it is important to clarify that although variables related to criminality were found to be important features in our HARM models to predict prospective physical aggression, most individuals with criminal histories do not pose an elevated

risk of violence. Moreover, these variables alone were insufficient to predict physical aggression in schizophrenia.

Across the 4 month and 12-month models, random forest appeared to outperform eight other algorithms, including other tree-based algorithms and linear models, in predicting physical aggression in patients over time. However, considering the overlapping confidence intervals between models, this cannot be determined definitively. As discussed, elsewhere [43], random forest tends to perform well with categorical variables, and can handle multicollinearity between highly correlated features. Similarly, models that incorporated boosting (boosted logistic regression, XGBoost, and AdaBoost) tended to perform well, as multicollinearity does not tend to be a significant issue as individual decision trees are used. Conversely, it is anticipated that lasso regression performed comparatively poorer than other algorithms across timepoints (balanced accuracy of 47.42-63.31), as it, unlike elastic net, lacks a sum of squared coefficient penalty term, which can help address multicollinearity. Additionally, extreme gradient boosting outperformed all other algorithms at 18-month follow-up (Balanced accuracy: 81.81%), although similar performance was observed using random forest (Balanced accuracy: 75.93%).

In contrast with existing actuarial tools, such as the VRAG and HCR-20, which consider a linear additive combination of variables to assess individual prospective risk, the HARM models incorporate a data-driven approach that allow for a non-linear weighting of importance between features, while also relying on theoretically sound and evidence-based risk factors, protective factors, and variables related to course of treatment. Moreover, the HARM models showed improvements in AUC relative to existing risk assessment tools, in predicting physical aggression at 4-month (AUC: 0.669-0.928), 12-month (AUC: 0.701-0.913), and 18-month (AUC: 0.597-0.870) follow-up. Additionally, the HARM models incorporate additional

performance measures, including sensitivity, specificity, balanced accuracy, overall model accuracy, as well as PPV and NPV, to better elucidate the goodness of fit of the models.

Furthermore, the data-driven HARM models may show utility in conjunction with clinician judgement of violent risk, to improve the accurate detection of patients with schizophrenia in forensic settings at risk of physical aggression. Overall, while the HARM models showed high NPVs (93.82-99.34) at 4-months follow-up, the PPVs were much lower (18.64-41.86) indicating a high degree of false positives, where many individuals who are classified as physically aggressive at 4-month follow-up, will fail to commit aggressive acts. Conversely, using clinical judgement alone at 4-months follow-up, although the PPVs were much higher (95.41-96.85), which illustrates a lower degree of false positives than the HARM models, the NPVs were notably lower, ranging from 18.08-63.16. Additionally, as shown in Supplementary Table S4, across clinical judgement models, model specificity was poor (57.14-61.50%). These results indicates that clinical judgement of violent risk performed little better than chance at identifying true negatives. As such, a high degree of false negatives is observed, where individuals who are physically aggressive at follow-up are incorrectly predicted to be non-aggressive. However, it is important to clarify that no statistically significant differences in error rates were observed between HARM and CLV models (McNemar's chi-square (χ^2) = 2.37, p=0.123).

Interestingly, as shown in Supplementary Table S5, a combined model incorporating both data-driven features and clinical judgement of violent risk did not show notable improvements in PPV, NPV, sensitivity, specificity, or AUC, although slightly higher balanced accuracy was observed (91.61% in a random forest model in the combined model, relative to 86.60% in the HARM model). Moreover, no statistically significant differences in error rates were observed between data-driven HARM models and those that incorporated clinician judgement of patient

risk (McNemar's χ^2= 0.5, p=0.479). While prospective validation is required, and a relatively small sample size was used in model development, machine learning models may show utility as an adjunct to clinical judgement to improve the accuracy of risk prediction for individualised care of patients with schizophrenia in forensic settings.

Limitations

The current study has some potential limitations. Although the study benefits from a longitudinal design, and showed similar variable importance across timepoints, a low base-rate of aggressive incidents was observed at 4-, 12-, and 18-month follow-up. As such, future studies with larger sample sizes will be required to determine the replicability of using evidence-based risk and protective factors, alongside treatment variables, to predict longitudinal physical aggression in patients with schizophrenia in forensic settings. Considering that the study used binary classification tasks, alongside baseline variables, to predict physical aggression, no hypothesis testing was performed, and as such, statistical power cannot be calculated. Since the present study had a low base rate of physical aggression, and relatively small sample size, it is possible that model accuracy is inflated.

Additionally, it is important to consider that these models were developed in a specific at-risk cohort of patients with schizophrenia who have a history of criminal offences. As such, these models may not be generalisable to detect aggressive behaviours in schizophrenia in general. Moreover, our models were developed largely using categorical features, which were transformed into binary variables using one-hot encoding. While several models were used that can handle multicollinearity, other methods, such as transforming features into principal

components [44], can be used to derive a set of uncorrelated variables. Additionally, further refinement is needed in prospective models, and a much smaller error rate is required to implement such predictive models as clinical tools. Similarly, it is important to consider the possibility that the present HARM models show artificially inflated AUC scores, even though similar performance was observed between the internally cross-validated and externally cross-validated models, across timepoints, due to a low base rate of physical aggression (n=26). As such, prospective validation with independent datasets is required to determine whether the HARM models show deflated scores within new samples. Nonetheless, these models are a notable improvement upon existing risk prediction in schizophrenia, which show a median AUC of 0.69 with an interquartile range of 0.60-0.77. Furthermore, it should be highlighted that within the HARM models the NPV/NPP substantially outperforms the PPV/PPP in the present analysis, which is related to the low base rate of physical aggression in the sample. Considering this, individuals who screen positive for non-aggression are more likely to show non-aggression at follow-up (true negatives), relative to individuals who screen positive for aggression (true positives).

Commented [DW1]: Update this description

Perspectives

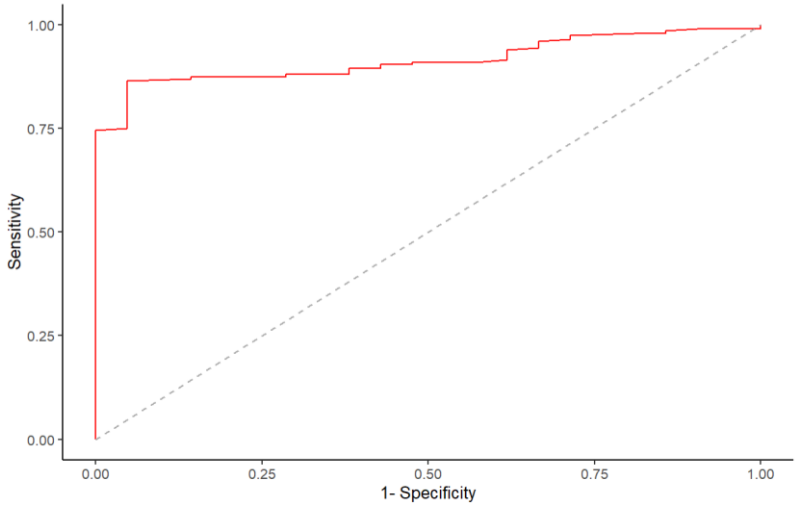
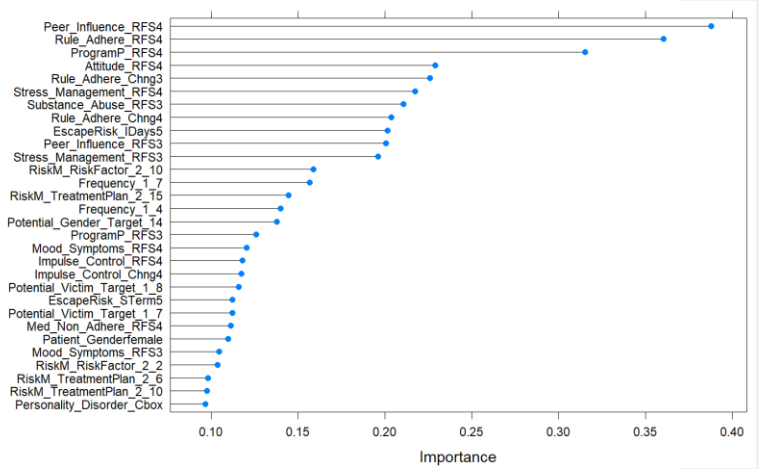
Moving forward, further refinement is required for individualised predictive models of physical aggression in patients with psychotic disorders. As detailed elsewhere [48], this may be facilitated by a wider framework when selecting input features in our models. Considering that model performance is directly dependent on the quality and quality of features, or variables, we have at our disposal – an exploratory data-driven approach to feature selection may be warranted. However, a hypothesis-driven framework is still required to evaluate the robustness and replicability of previously identified predictors. Our models identified several known risk

factors, including changes in impulsivity, attitude, psychotic symptoms, and mood symptoms as important predictors of longitudinal physical aggression in patients with schizophrenia. Considering this, targeting modifiable risk factors, including poor program participation, mood symptoms, and improving impulse control, may be useful strategies to curtail physical aggression in patients with schizophrenia in forensic settings.

Prospective work may benefit from the inclusion of additional psychometric scales pertaining to these risk factors, to better elucidate more subtle changes in attitudes and behaviour that precede physical aggression in high-risk individuals with schizophrenia. As such, future studies may benefit from including these variables as candidate features, alongside other presumed risk, and protective factors. Additionally, future studies may benefit from incorporating large-scale electronic health record (EHR) data, to both identify more time-dependent predictors with greater granularity, as well as potentially identify adjunctive medications that may decrease the risk of aggression in individuals, thereby serving as a repurposing candidate for prospective trials.

Moreover, the current models utilise categorical features, and prospective studies may benefit from incorporating numeric variables that may better capture nuances in factors such as impulsivity, rule adherence, attitude, mood symptoms, and psychotic symptoms. While previously identified risk factors are important to include in our models to assess the replicability of these effects, novel markers are also required to improve our understanding of the mechanisms underlying physical aggression in patients with schizophrenia. For instance, it may be warranted to include routinely collected sensor data, such as blood pressure and heart rate variability, wearables such as actigraphy, and blood biomarkers in prospective models to identify novel markers that may improve the performance of models predicting aggression in patients

with psychotic disorders. Other modalities, such as neuroimaging and electroencephalography may also be useful when combined with structured and unstructured clinical data.

A)  	B) 
<u>Overall model (67 variables)</u> Balanced Accuracy = 86.60% Accuracy = 87.33% (95% CI: 82.21-91.41) Sensitivity = 85.71% Specificity = 87.50% PPV: 41.86 NPV: 98.31	<u>Embedded feature selection (23 variables)</u> Balanced Accuracy = 89.36% Accuracy = 84.62% (95% CI: 79.17, 89.10) Sensitivity = 95.23% Specificity = 83.50% PPV: 37.73 NPV: 99.40

AUC = 0.914 (95% CI: 0.872-0.951) Positive Class: Physical Aggression	AUC = 0.949 (95% CI: 0.914-0.975) Positive Class: Physical Aggression

Figure 1 - AUC, Variable Importance, and Model Performance at 4-month follow-up

The best performance in predicting physical aggression at 4-month follow-up in patients with schizophrenia in forensic settings was observed using a random forest model. ROC curves were generated using the *roc* function in R, as depicted in Supplementary Figure S5a. 95% CI of AUC was calculated using the *ci.auc* function in the pROC package in R, with 5000 stratified bootstrap replicates. A variable importance plot was generated using the varImp function in the caret package in R, showcasing the top 23 features. Model performance is shown for both the total model comprising 67 variables, and a model comprising the top 30 important features within a random forest model (as shown in Figure S2a), determined using the total decrease in node impurity, calculated using the Gini Index, from splitting on the variable, averaged over all trees. Important features in the random forest model included worsening peer influence (Peer_Influence_RFS4), worsening rule adherence (Rule_Adhere_RFS4), poor program participation (ProgramP_RFS4), worsening attitude (Attitude_RFS4), worsening stress management (Stress_Management_RFS4), and changes in substance abuse (Substance_Abuse_RFS3). Similarly, frequency of treatment (Frequency_1_7, Frequency_1_4), worsening mood symptoms (Mood_Symptoms_RFS4), changes in impulse control (Impulse_Control_RFS4, Impulse_Control_Chng4), worsening medication non-adherence (Med_Non_Adhere_RFS4), and personality disorder (Personality_Disorder_Cbox) were among the important features in the model. Further details regarding candidate features can be found in Supplementary Table S3.

	Nonaggressive (n=170)	**Aggressive (n=26)**	***p*-Value**
Age (years) [a]	41.75±13.12	37.53 ±11.87	0.1777
Gender [b]			
Male	149 (87.6%)	19 (73.0%)	0.0480 (*)
Female	21 (12.3%)	7 (26.9%)	
Index Offences			0.8026
Attempt Murder, Assault & Related Offences	104 (80.6%	15 (75%)	
Escape Custody	45 (26.4%)	10 (38.4%)	
Weapon Related Offence			
Arson	34 (20.0%)	6(23.0%)	
Mischief, Driving-Related, and Miscellaneous Offences	20 (11.7%)	2 (7.6%)	
	79 (61.2%)	14 (70%)	
History of Substance Abuse	148 (87.05%)	23 (88.46 %)	0.8417
Personality Disorder	38 (22.35%)	13 (50.00%)	0.027 (*)

Table 1 - Demographics

Demographic and clinical characteristics of patients with schizophrenia at 4-month follow-up (n =151). A one-way ANOVA was used for numeric variables, and data are given as mean and standard deviation. A Chi-Square test with Yates correction, with a significance level of 0.5, was performed for categorical variables, including Gender, Index Offences, history of substance abuse, and diagnosis of a personality disorder. A statistically significant difference was observed between aggressive and non-aggressive patients with respect to diagnosis of a personality disorder ($X2 = 8.95$, $p = .002$), and gender ($X2 = 3.90$, $p = 0.04$). No significant group differences were observed with respect to age ($p = 0.17$), intake offence ($p = 0.80$), or history of substance abuse ($p = 0.84$).

Boosted Logistic Regression	**Elastic Net**	**Lasso Regression**
Balanced Accuracy = 84.16%	Balanced Accuracy = 82.22%	Balanced Accuracy = 62.83%
Accuracy = 74.20%	Accuracy = 83.26%	Accuracy = 59.73%
(95% CI: 78.67-88.71)	(95% CI: 77.67-87.93)	(95% CI: 52.94-66.25)
Sensitivity = 61.90%	Sensitivity = 80.95%	Sensitivity = 66.66%
Specificity = 86.50%	Specificity = 83.50%	Specificity = 59.00%
PPV: 32.50	PPV: 34.00	PPV: 14.58
NPV: 95.58	NPV: 97.66	NPV: 94.40
AUC: 0.903	AUC: 0.815	AUC: 0.712
(95% CI: 0.858-0.942	(95% CI: 0.656-0.947)	(95% CI: 0.584-0.833)
kNN	**AdaBoost**	**XGBoost**
Balanced Accuracy = 64.19%	Balanced Accuracy = 74.83%	Balanced Accuracy = 86.25%
Accuracy = 73.76%	Accuracy = 81.45	Accuracy = 82.81%
(95% CI: 67.43-79.43)	(95% CI: 75.69-86.35)	(95% CI: 77.17-87.54)
Sensitivity = 52.38%	Sensitivity = 66.66%	Sensitivity = 90.47%
Specificity = 76.00%	Specificity = 83.00%	Specificity = 82.00%
PPV: 18.64	PPV: 29.16	PPV: 34.54
NPV: 93.82	NPV: 95.95	NPV: 98.79
AUC: 0.669	AUC: 0.826	AUC: 0.928
(95% CI: 0.551-0.784)	(95% CI: 0.764-0.883)	(95% CI: 0.885-0.963)
Random Forest	**Bagged CART**	**Conditional Forest**
Balanced Accuracy = **86.60**%	Balanced Accuracy = 74.21%	Balanced Accuracy = 85.36%
Accuracy = 87.33%	Accuracy = 76.47%	Accuracy = 77.38%
(95% CI: 82.21-91.41)	(95% CI: 70.32-81.90)	(95% CI: 71.28-82.72)
Sensitivity = 85.71%	Sensitivity = 71.42%	Sensitivity = 95.23%
Specificity = 87.50%	Specificity = 77.00%	Specificity = 75.50%
PPV: 41.86	PPV: 24.59	PPV: 28.98
NPV: 98.31	NPV: 96.50	NPV: 99.34

AUC: 0.914 (95% CI: 0.872-0.951)	AUC: 0.928 (95% CI: 0.886-0.964)	AUC: 0.914 (95% CI: 0.869-0.953)

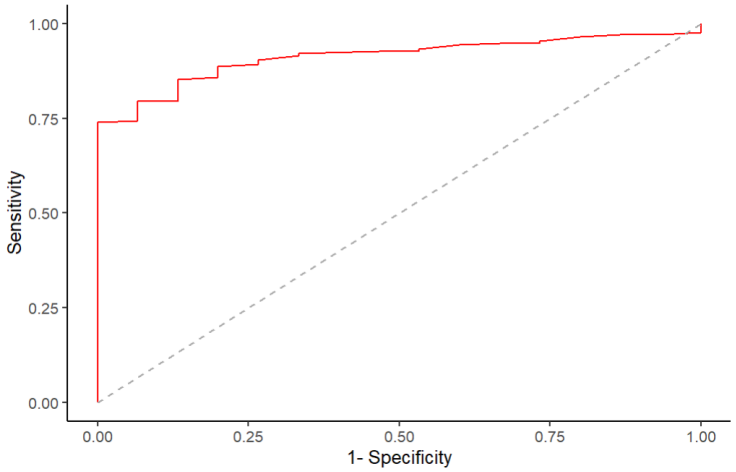
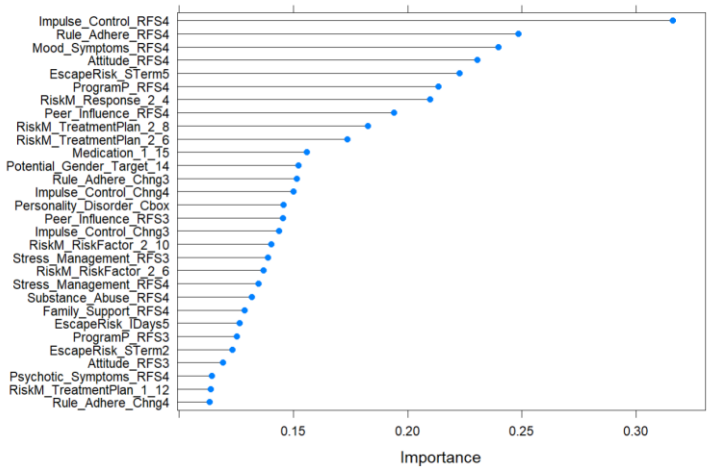
Table 2 - Model Performance - 4-month follow-up

Model performance in predicting prospective physical aggression in patients with schizophrenia in the testing dataset (40%), using baseline risk factors, protective factors, and treatment status. Seventeen instances of aggression were recorded at baseline, and twenty-one instances were recorded at 4-month follow-up, respectively. Baseline assessments involved the first five Hamilton Anatomy of Risk Management (HARM) clinical evaluations, and follow-up involved assessments 10-14, corresponding to 4-month follow-up. As such, patients could commit more than one aggressive incident during baseline, and follow-up, with each instance recorded as a separate event. Across binary classification models, aggression was considered as the positive class, and non-aggression as the negative class, respectively. The best performance was observed using random forest, followed by conditional forest, elastic net, XGBoost, and boosted logistic regression. Across most models, the true positives (sensitivities) were higher than true negatives (specificities), suggests that the models performed better in discriminating those with physical aggression, relative to non-aggression. However, considering the low base rate of physical aggression in the overall sample (15.29%) the negative predictive values (PPV), were much higher than the positive predictive values (NPV), indicating a much higher ratio of true negative predictions (non-aggression), considering all positive predictions, across models.

Level	Incident
9	Critical incident - possible life and death - possible call police
8	Violent unprovoked assault
7	Violent assault
6	Push/shove
5	Destruction of property
4	Improper physical contact
3	Intimidating, threatening, personal space violated
2	Intimidating, raised voice
1	Rude, argumentative

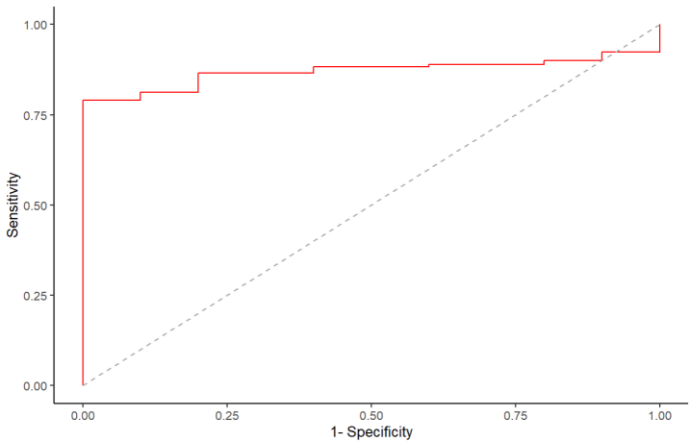
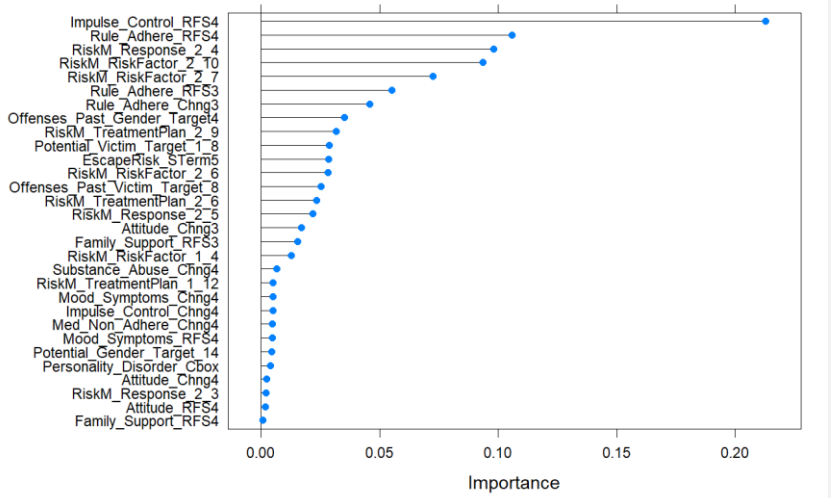
Supplementary Figure S1 - Aggressive Incidents Scale

The AIS provides a standardised method to longitudinally record aggressive incidents in patients within forensic settings. In the current study, physical aggressive outcomes at follow-up time-points were dichotomized (yes/no) according to whether an AIS score of ≥4 was observed. Individuals with schizophrenia who showcased an AIS score ≤3 at follow-up, were considered non physically aggressive.

<table>
<tr>
<td>A)
 
</td>
<td>B)
 
</td>
</tr>
<tr>
<td><u>Overall model (67 variables)</u>
Balanced Accuracy = 81.81%
Accuracy = 83.43%
(95% CI: 77.19-88.53)
Sensitivity = 80.00%
Specificity = 83.62%
PPV: 22.22
NPV: 98.62
AUC: 0.893 (95% CI: 0.843-0.936)</td>
<td><u>Embedded feature selection (23 variables)</u>
Balanced Accuracy = 84.24%
Accuracy = 82.20%
(95% CI: 76.02-87.35)
Sensitivity = 86.66%
Specificity = 81.81%
PPV: 28.88
NPV: 98.63
AUC: 0.868 (95% CI: 0.791-0.932)</td>
</tr>
</table>

Supplementary Figure S2 - AUC, Variable Importance, and Model Performance at 12-month follow-up

The best performance in predicting physical aggression at 12-month follow-up in patients with schizophrenia in forensic settings was observed using a random forest model. ROC curves were generated using the *roc* function in R, as depicted in Supplementary Figure S5a. 95% CI of AUC was calculated using the *ci.auc* function in the pROC package in R, with 5000 stratified bootstrap replicates. A variable importance plot was generated using the varImp function in the caret package in R, showcasing the top 23 features. Model performance is shown for both the total model comprising 67 variables, and a model comprising the top 30 important features within a random forest model (as shown in Figure S2a), determined using the total decrease in node impurity, calculated using the Gini Index, from splitting on the variable, averaged over all trees. Important baseline features in the random forest model included changes in impulse control (Impulse_Control_RFS4, Impulse_Control_Chng4, Impulse_Control_Chng3), changes in rule adherence (Rule_Adhere_RFS4, Rule_Adhere_Chng3, Rule_Adhere_Chng4), worsening mood symptoms (Mood_Symptoms_RFS4), worsening attitude (Attitude_RFS4), short-term escape risk (EscapeRisk_STerm5), worsening program participation (ProgramP_RFS4), high engagement in a treatment program (RiskM_Response_2_4), worsening peer influence (Peer_Influence_RFS4), enrolment in a psychoeducational or recreational program (RiskM_TreatmentPlan_2_8, RiskMTreatmentPlan_2_6), and current use of Haloperidol (Medication_1_15). Other important features in the model included the presence of a personality disorder (Personality_Disorder_Cbox), changes in stress management (Stress_Management_RFS3, StressManagement_RFS4), worsening substance abuse (Substance_Abuse_RFS4), worsening family support (Family_Support_RFS4), and worsening psychotic symptoms (Psychotic_Symptoms_RFS4), and current enrolment in family support therapy (RiskM_TreatmentPlan_1_12). Further details regarding candidate features can be found in Supplementary Table S3.

A)  	B) 
<u>Overall model (67 variables)</u> Accuracy = 83.43% (95% CI: 77.19-88.53) Balanced Accuracy = 81.81% Sensitivity = 80.00% Specificity = 83.62% PPV = 22.22 NPV = 98.62 AUC = 0.870 (95% CI: 0.814-0.918)	<u>Embedded feature selection (23 variables)</u> Accuracy = 70.71% (95% CI: 64.50-76.40) Balanced Accuracy = 66.39% Sensitivity = 60.00% Specificity = 85.38% PPV = 19.35 NPV = 97.33 AUC = 0.898 (95% CI: 0.825-0.956)

Supplementary Figure S3 - AUC, Variable Importance, and Model Performance at 18-month follow-up

The best performance in predicting physical aggression at 18-month follow-up in patients with schizophrenia in forensic settings was observed using an Extreme Gradient Boosting (XGBoost) model. ROC curves were generated using the *roc* function in R, as depicted in Supplementary Figure S5a. 95% CI of AUC was calculated using the *ci.auc* function in the pROC package in R, with 5000 stratified bootstrap replicates. A variable importance plot was generated using the varImp function in the caret package in R, showcasing the top 23 features. Among them, worsening impulse control (Impulse_Control_RFS4), worsening rule adherence (Rule_Adhere_RFS4), high/moderate engagement in an existing treatment program/intervention (RiskM_Response_2_4, RiskM_Response_5), stress management (RiskM_TreatmentPlan_2_9), changes in family support (Family_Support_RFS3), changes in rule adherence (Rule_Adhere_RFS4), treatment with individual psychotherapy (RiskM_TreatmentPlan_2_9), changes in mood symptoms (Mood_Symptoms_Chng4, Mood_Symptoms_Chng4), medication non-adherence (Med_Non_Adhere_Chng4), and a history of a personality disorder (Personality_Disorder_Cbox) were among the most important features in the model. Further details regarding candidate features can be found in Supplementary Table S3.

Boosted Logistic Regression	**Elastic Net**	**Lasso Regression**
Balanced Accuracy = 79.70% Accuracy = 68.31% (95% CI: 61.03-74.97) Sensitivity = 93.33% Specificity = 66.07% PPV = 19.71 NPV = 99.10 AUC = 0.865 (95% CI: 0.772-0.935)	Balanced Accuracy = 75.01% Accuracy = 76.44% (95% CI: 69.77-82.27) Sensitivity = 73.33% Specificity = 76.05% PPV = 21.15 NPV = 97.12 AUC = 0.794 (95% CI: 0.704-0.876)	Balanced Accuracy = 63.31% Accuracy = 71.23% (95% CI: 64.77-77.99) Sensitivity = 53.33% Specificity = 73.29% PPV = 14.54 NPV = 94.85 AUC = 0.721 (95% CI: 0.589-0.841)
kNN	**AdaBoost**	**XGBoost**
Balanced Accuracy = 65.58% Accuracy = 75.92% (95% CI: 69.21-81.80) Sensitivity = 53.33% Specificity = 77.84% PPV = 17.02 NPV = 95.13 AUC = 0.701 (95% CI: 0.579-0.816)	Balanced Accuracy = 79.48% Accuracy = 79.06% (95% CI: 0.725-0.846) Sensitivity = 80.00% Specificity = 78.97% PPV = 24.49 NPV = 97.88 AUC = 0.883 (95% CI: 0.825-0.934)	Balanced Accuracy = 83.88% Accuracy = 75.92% (95% CI: 69.21-81.80) Sensitivity = 93.33% Specificity = 74.44% PPV = 23.72 NPV = 99.24 AUC = 0.841 (95% CI: 0.711-0.931)
Random Forest	**Bagged CART**	**Conditional Forest**
Balanced Accuracy = **86.15%** Accuracy = 80.10% (95% CI: 73.73-85.52) Sensitivity = 93.33% Specificity = 78.97% PPV = 27.45 NPV = 99.28 AUC = 0.911 (95% CI: 0.862-0.954)	Balanced Accuracy = 81.17% Accuracy = 76.44% (95% CI: 69.77-82.27) Sensitivity = 86.66% Specificity = 75.56% PPV = 23.21 NPV = 98.51 AUC = 0.858 (95% CI: 0.798-0.912)	Balanced Accuracy = 82.17% Accuracy = 72.77% (95% CI: 65.88-78,95) Sensitivity = 93.33% Specificity = 71.02% PPV = 21.53 NPV = 99.20 AUC = 0.913 (95% CI: 0.859-0.955)

Supplementary Table S1 - Model Performance (12-month follow-up)

Model performance in predicting prospective physical aggression in patients with schizophrenia in the testing dataset (40%), using baseline risk factors, protective factors, and treatment status. The best performance was observed using random forest, followed by XGBoost, conditional forest, boosted logistic regression, adaboost, and elastic net. The positive class corresponded to physical-aggression and the negative class corresponded to non-aggression, respectively. Across most models, the true positives (sensitivities) were higher than true negatives (specificities), suggests that the models performed better in discriminating those with physical aggression, relative to non-aggression. However, considering the low base rate of physical aggression at 12-months (8.28%) the positive predictive values (PPV), were much lower than the negative predictive values (NPV), indicating a much higher ratio of true negative predictions (non-aggression), considering all positive predictions, across models.

Boosted Logistic Regression	**Elastic Net**	**Lasso Regression**
Balanced Accuracy = 54.44% Accuracy = 85.08% (95% CI: 79.04-89.93) Sensitivity = 40.00% Specificity = 78.36% PPV = 9.75 NPV = 95.71 AUC = 0.732 (95% CI: 0.614-0.843)	Balanced Accuracy = 56.52% Accuracy = 80.11% (95% CI: 73.54-85.66) Sensitivity = 30.00% Specificity = 83.04% PPV = 9.37 NPV = 95.30 AUC = 0.611 (95% CI: 0.381-0.813)	Balanced Accuracy = 47.42% Accuracy = 71.82% (95% CI: 64.67-78.25) Sensitivity = 20.00% Specificity = 74.85% PPV = 4.44 NPV = 94.11 AUC = 0.628 (95% CI: 0.528-0.730)
kNN	**AdaBoost**	**XGBoost**
Balanced Accuracy = 46.22% Accuracy = 77.90% (95% CI: 71.15-83.72) Sensitivity = 40.00% Specificity = 80.11% PPV = 10.52 NPV = 95.80 AUC = 0.597 (95% CI: 0.717-0.842)	Balanced Accuracy = 59.47% Accuracy = 76.80% (95% CI: 69.96-82.74) Sensitivity = 40.00% Specificity = 78.94% PPV = 10.00 NPV = 95.74 AUC = 0.747 (95% CI: 0.654-0.839)	**Balanced Accuracy = 81.81%** Accuracy = 83.43% (95% CI: 77.19-88.53) Sensitivity = 80.00% Specificity = 83.62% PPV = 22.22 NPV = 98.62 AUC = 0.870 (95% CI: 0.814-0.918)
Random Forest	**Bagged CART**	**Conditional Forest**

Balanced Accuracy = 75.93% Accuracy = 81.22% (95% CI: 0.747-0.866) Sensitivity = 70.00% Specificity = 81.87% PPV = 18.42 NPV = 97.90 AUC = 0.868 (95% CI: 0.797-0.928)	Balanced Accuracy = 65.50% Accuracy = 72.38% (95% CI: 65.25-78.75) Sensitivity = 60.00% Specificity = 73.09% PPV = 11.53 NPV = 96.89 AUC = 0.750 (95% CI: 0.652-0.785)	Balanced Accuracy = 63.27% Accuracy = 83.98% (95% CI: 77.81-89.00) Sensitivity = 40.00% Specificity = 86.55% PPV = 14.81 NPV = 96.10 AUC = 0.852 (95% CI: 0.775-0.924)

Supplementary Table S2 - Model Performance (18-month follow-up)

Model performance reported for predicting prospective physical aggression in patients with schizophrenia in the testing set (40%), using baseline risk factors, protective factors, and treatment status variables. The highest balanced accuracy and AUC was observed within an XGBoost model, with similar performance using random forests. On average, model sensitivities (true positives) were more variable than across 6-months and 12-months timepoints, ranging from 20-80%. Across most models, the true negatives (specificities) were higher than true positives (sensitivities), suggests that the models performed better in discriminating those with physical aggression, relative to non-aggression. However, considering the low base rate of physical aggression (7.57%) the negative predictive values (PPV), were much lower than the negative predictive values (NPV), indicating a much higher ratio of true negative predictions (non-aggression), considering all positive predictions, across models.

Variable	**Description**
Patient_Gender	Patient Gender
Arson_IO_Cbox	Arson at intake offence
Assaults_IO_Cbox	Assault at intake offence
Homicide_IO_Cbox	Homicide at intake offence
Kidnapping_IO_Cbox	Kidnapping at intake offence
Robbery_IO_Cbox	Robbery at intake offence
Driving_IO_Cbox	Driving offence at intake

Sexual_CC_Cbox	Sexual offence at intake
Frauds_CC_Cbox	Fraud offence at intake
Offenses_Past_Gender_Target	Past gender target of patient
Substance_Use_Cbox	Substance use at intake
Cognitive_Deficits_Cbox	Cognitive deficits at intake
Other1_HistRiskFactor_Cbox	Other historical risk factor of patient
Other2_HistRiskFactor_Cbox	Other historical risk factor of patient
Mood_Symptoms_Cbox	Mood symptoms (historical risk factor)
Impulse_Control_Cbox	Impulse control (historical risk factor)
ProgramP_Cbox	Program participation (risk factor)
Substance_Abuse_Cbox	Substance abuse (risk factor)
Med_Non_Adhere_Cbox	Medication non-adherence (risk factor)
Attitude_Cbox	Attitude (risk factor)
Stress_Management_Cbox	Stress management (risk factor)
Anger_Management_Cbox	Anger management (risk factor)
Peer_Influence_Cbox	Peer influence (risk factor)
Other_RiskFactor_Cbox	Other risk factors
Rule_Adhere_RFS	Rule Adherence Risk Factor Status (Managed, monitor, needs improvement)
Criminal_Harassment_IO_Cbox	Criminal Harassment, Utter threats, and related

	offences at intake
Insight_Ill_RFS	Insight into illness risk factor status (Managed, monitor, needs improvement)
Mood_Symptoms_RFS	Mood symptoms (Managed, monitor, needs improvement)
Psychotic_Symptoms_RFS	Psychotic symptoms (Managed, monitor, needs improvement)
Impulse_Control_RFS	Impulse control (Managed, monitor, needs improvement)
ProgramP_RFS	Program participation (Managed, monitor, needs improvement)
Substance_Abuse_RFS	Substance abuse (Managed, monitor, needs improvement)
Med_Non_Adhere_RFS	Medication non-adherence (Managed, monitor, needs improvement)
Attitude_RFS	Attitude (Managed, monitor, needs improvement)
Stress_Management_RFS	Stress management (Managed, monitor, needs improvement)
Family_Support_RFS	Family support (Managed, monitor, needs improvement)
Peer_Influence_RFS	Peer influence (Managed, monitor, needs improvement)

Rule_Adhere_Chng	Rule adherence (Better, worse, same)
Insight_Ill_Chng	Insight into illness (Better, worse, same)
Mood_Symptoms_Chng	Mood symptoms (Better, worse, same)
Psychotic_Symptoms_Chng	Psychotic symptoms (Better, worse, same)
Impulse_Control_Chng	Impulse control (Better, worse, same)
ProgramP_Chng	Program participation (Better, worse, same)
Substance_Abuse_Chng	Substance abuse (Better, worse, same)
Med_Non_Adhere_Chng	Medication non-adherence (Better, worse, same)
Attitude_Chng	Attitude (Better, worse, same)
Stress_Management_Chng	Stress Management (Better, worse, same)
Family_Support_Chng	Family Support (Better, worse, same)
Peer_Influence_Chng	Peer influence

	(Better, worse, same)
Potential_Gender_Target	Anticipated gender of potential victim
EscapeRisk_IDays	Escape risk immediate (days)
EscapeRisk_STerm	Escape risk short-term
Medication	Medications
Frequency	Frequency of medications
Class_Of_Medication	Class of medications
Offenses_Past_Victim_Target	Gender of prior victim
Protective_Factors_1	Protective factors: employment, leisure activities, financial stability, motivation for treatment, positive attitude, realistic goals, stable intimate relationship, stable housing, external control, positive social support, none
Protective_Factors_2	Protective factors: employment, leisure activities, financial stability, motivation for treatment, positive attitude, realistic goals, stable intimate relationship, stable housing, external control, positive social support, none
Potential_Behaviours	Anticipated behaviors - physical aggression, arson, criminal harassment, extreme property damage, robbery, sexual aggression/behavior, terrorism, verbal aggression
Potential_Victim_Target	Potential target of subsequent criminal offences - staff, known persons, children, stranger, acquaintance, family member, indiscriminate, serious property damage
RiskM_RiskFactor_1	Risk factors: rule adherence, insight into illness, mood symptoms, psychotic symptoms, impulse control, program participation, substance abuse, med

	non-adherence, attitude/cooperation, stress management, anger management, family support, peer influence
RiskM_RiskFactor_2	Risk factors: rule adherence, insight into illness, mood symptoms, psychotic symptoms, impulse control, program participation, substance abuse, med non-adherence, attitude/cooperation, stress management, anger management, family support, peer influence
RiskM_TreatmentPlan_1	Treatment plan: substance abuse program, anger management, social skills training, mindfulness/relaxation, stress management, recreational program, vocational program, psychoeducation, individual psychotherapy, group therapy, medication, spiritual support, discharge planning, behavioural therapy, dialectical behavioural therapy, occupational therapy
RiskM_TreatmentPlan_2	Treatment plan: substance abuse program, anger management, social skills training, mindfulness/relaxation, stress management, recreational program, vocational program, psychoeducation, individual psychotherapy, group therapy, medication, spiritual support, discharge planning, behavioural therapy, dialectical behavioural therapy, occupational therapy
RiskM_Response_1	Patient response: referral pending, declined participation, on waitlist, highly engaged, moderately engaged, low engagement, sporadic attendance, disruptive in program, withdrew from program, completed, expelled from program, medication adherent, medication non-adherent
RiskM_Response_2	Patient response: referral pending, declined participation, on waitlist, highly engaged, moderately engaged, low engagement, sporadic attendance, disruptive in program, withdrew from program, completed, expelled from program, medication adherent, medication non-adherent

LTRE	Long-term risk assessment

Supplementary Table S3 - List of Candidate Features and eHARM Measures

Baseline risk factors collected over three baseline assessments were used to predict subsequent physical aggression at 4-month, 12 months, and 18-month follow-ups. Only variables included in the list above were used as candidate features in model development.

Random Forest	Elastic Net	Conditional Forest
Balanced Accuracy = 71.26% Accuracy = 63.35% (95% CI: 56.62, 69.71) Sensitivity = 80.95% Specificity = 61.50% PPV = 96.85 NPV = 18.08 AUC = 0.800 (95% CI: 0.646-0.930)	Balanced Accuracy = 75.32% Accuracy = 90.05% (95% CI: 85.32, 93.66) Sensitivity = 93.50% Specificity = 57.14% PPV = 95.41 NPV = 48.00 AUC = 0.893 (95% CI: 0.826-0.948)	Balanced Accuracy = **76.82%** Accuracy = 92.76% (95% CI: 88.51, 95.81) Sensitivity = 96.50% Specificity = 57.14% PPV = 95.54 NPV = 63.16 AUC = 0.861 (95% CI: 0.784-0.927)

Supplementary Table S4: Clinician-rated clinical-likelihood of violence model

A summary of the top performing algorithms, according to balanced accuracy and AUC. Across models, balanced accuracy ranged from 71.26-76.82%, with the highest balanced accuracy in a conditional forest model. While variation was observed in sensitivity and specificity across models, the number of true positives (sensitivity) of physical aggression was higher, relative to true negatives (specificity) of non-aggression. Therefore, in the current sample, clinical judgement alone showed a high detection rate of actual instances of physical aggression. However, models performed little better than chance in identifying true negatives. As such, clinical judgement shows a high level of false negatives, where individuals who are physically aggressive at follow-up are incorrectly predicted to be non-aggressive.

Random Forest	Conditional Forest	XGBoost
Balanced Accuracy = **91.61%** Accuracy = 88.69% (95% CI: 83.76, 92.54) Sensitivity = 95.23% Specificity = 88.00% PPV = 45.45 NPV = 99.43 AUC = 0.945 (95% CI: 0.907-0.974)	Balanced Accuracy = 85.61% Accuracy = 77.83% (95% CI: 71.77, 83.12) Sensitivity = 95.23% Specificity = 76.00% PPV = 29.41 NPV = 99.34 AUC = 0.934 (95% CI: 0.894-0.967)	Balanced Accuracy = 82.73% Accuracy = 75.11% (95% CI: 68.87, 80.67) Sensitivity = 95.23% Specificity = 73.00% PPV = 27.02 NPV = 99.32 AUC = 0.919 (95% CI: 0.873-0.958)

Supplementary Table S5: Combined model of HARM features and clinician-rated clinical-likelihood of violence model

A summary of the top performing algorithms, according to balanced accuracy and AUC. Across models, balanced accuracy ranged from 68.31-91.61%, with the highest balanced accuracy and AUC in a random forest model. A statistically significant difference was observed in classifier

performance between a combined model, which incorporated both HARM features and clinician rated clinical likelihood of violence (CLV), and CLV alone (McNemar's $\chi 2 = 10.22$, $p = 0.001$).

Declaration of competing interest:

Devon Watts reports a CIHR Doctoral Scholarship, outside of the submitted work. Taiane de Azevedo Cardoso reports a CIHR Postdoctoral Scholarship, outside of the submitted work. Heather Moulden, Mini Mamak, Casey Upfold, and Gary Chaimowitz report no biomedical financial interests or potential conflicts of interest. Flávio Kapczinski reports personal fees from Daiichi sankyo, and Janssen-Cilag; grants from Stanley Medical Research Institute 07TGF/1148, grants from INCT - CNPq 465458/2014-9, and from the Canadian Foundation for Innovation - CFI, outside the submitted work.

References

1. Faay, M. D. M. & Sommer, I. E. Risk and Prevention of Aggression in Patients with Psychotic Disorders. *American Journal of Psychiatry* **178**, 218–220 (2021).
2. Fazel, S., Wolf, A., Palm, C. & Lichtenstein, P. Violent crime, suicide, and premature mortality in patients with schizophrenia and related disorders: a 38-year total population study in Sweden. *The Lancet Psychiatry* **1**, 44–54 (2014).
3. Whiting, D., Gulati, G., Geddes, J. R. & Fazel, S. Association of Schizophrenia Spectrum Disorders and Violence Perpetration in Adults and Adolescents From 15 Countries. *JAMA Psychiatry* **79**, 120 (2022).
4. Douglas, K. S., Guy, L. S. & Hart, S. D. Psychosis as a risk factor for violence to others: A meta-analysis. *Psychological Bulletin* **135**, 679–706 (2009).
5. Penn, D. L., Kommana, S., Mansfield, M. & Link, B. G. Dispelling the Stigma of Schizophrenia: II. The Impact of Information on Dangerousness. *Schizophrenia Bulletin* **25**, 437–446 (1999).
6. Buchanan, A., Sint, K., Swanson, J. & Rosenheck, R. Correlates of Future Violence in People Being Treated for Schizophrenia. *American Journal of Psychiatry* **176**, 694–701 (2019).
7. Moulin, V. *et al.* Impulsivity in early psychosis: A complex link with violent behaviour and a target for intervention. *European Psychiatry* **49**, 30–36 (2018).
8. Storvestre, G. B. *et al.* Childhood Trauma in Persons With Schizophrenia and a History of Interpersonal Violence. *Frontiers in Psychiatry* **11**, (2020).
9. Keers, R., Ullrich, S., DeStavola, B. L. & Coid, J. W. Association of Violence With Emergence of Persecutory Delusions in Untreated Schizophrenia. *American Journal of Psychiatry* **171**, 332–339 (2014).
10. Swanson, J. W. *et al.* A National Study of Violent Behavior in Persons With Schizophrenia. *Archives of General Psychiatry* **63**, 490 (2006).
11. de Girolamo, G. *et al.* A multinational case−control study comparing forensic and non-forensic patients with schizophrenia spectrum disorders: the EU-VIORMED project. *Psychological Medicine* 1–11 (2021) doi:10.1017/S0033291721003433.
12. Whiting, D., Lichtenstein, P. & Fazel, S. Violence and mental disorders: a structured review of associations by individual diagnoses, risk factors, and risk assessment. *The Lancet Psychiatry* **8**, 150–161 (2021).
13. Fazel, S., Wolf, A., Palm, C. & Lichtenstein, P. Violent crime, suicide, and premature mortality in patients with schizophrenia and related disorders: a 38-year total population study in Sweden. *The Lancet Psychiatry* **1**, 44–54 (2014).

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

14. Sariaslan, A., Larsson, H. & Fazel, S. Genetic and environmental determinants of violence risk in psychotic disorders: a multivariate quantitative genetic study of 1.8 million Swedish twins and siblings. *Molecular Psychiatry* **21**, 1251–1256 (2016).

15. Fleischman, A., Werbeloff, N., Yoffe, R., Davidson, M. & Weiser, M. Schizophrenia and violent crime: a population-based study. *Psychological Medicine* **44**, 3051–3057 (2014).

16. Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. Criminal Justice and Behavior, 37(9), 965–988. https://doi.org/10.1177/0093854810374274

17. Kröner, C., Stadtland, C., Eidt, M. and Nedopil, N., 2007. The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. Criminal Behaviour and Mental Health, 17(2), pp.89-100.

18. Singh, J.P., Serper, M., Reinharth, J. and Fazel, S., 2011. Structured assessment of violence risk in schizophrenia and other psychiatric disorders: a systematic review of the validity, reliability, and item content of 10 available instruments. Schizophrenia bulletin, 37(5), pp.899-912.

19. Michel, S.F., Riaz, M., Webster, C., Hart, S.D., Levander, S., Müller-Isberner, R., Tiihonen, J., Repo-Tiihonen, E., Tuninger, E. and Hodgins, S., 2013. Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and dynamic risk factors. International Journal of Forensic Mental Health, 12(1), pp.1-13.

20. Lobo, J.M., Jiménez-Valverde, A. and Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), pp.145-151.

21 Wiens, J. & Shenoy, E. S. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases* **66**, 149–153 (2018).

22. Cearns, M., Hahn, T. & Baune, B. T. Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry* **9**, 271 (2019).

23. *DIAGNOSTIC AND STATISTICAL MANUAL OF DSM-5 ™*.

24. Mullally, K., Mamak, M. & Chaimowitz, G. A. The next generation of risk assessment and management. *International Journal of Risk and Recovery* **1**, 21–26 (2018).

25. Cook, A. N. *et al*. Validating the Hamilton Anatomy of Risk Management–Forensic Version and the Aggressive Incidents Scale. *Assessment* **25**, 432–445 (2018).

26. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* (2008) doi:10.18637/jss.v028.i05.

27. Kuhn, M. caret Package. *Journal Of Statistical Software* (2008).

28. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* (2002).

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

29. Schonlau RAND, M. *Boosted regression (boosting): An introductory tutorial and a Stata plugin*. *The Stata Journal* vol. 5 (2005).

30. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320 (2005).

31. Zhang, Z. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine* **4**, 218–218 (2016).

32. Freund, Y. & Schapire, R. E. *A Short Introduction to Boosting*. *Journal of Japanese Society for Artificial Intelligence* vol. 14 www.research.att.com/ (1999).

33. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. (2016) doi:10.1145/2939672.2939785.

34. Breiman, L. Random forests. *Machine Learning* (2001) doi:10.1023/A:1010933404324.

35. Sutton, C. D. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics* vol. 24 303–329 Preprint at https://doi.org/10.1016/S0169-7161(04)24011-1 (2005).

36. Nasejje, J. B., Mwambi, H., Dheda, K. & Lesosky, M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology* **17**, 115 (2017).

37. Dash, M. & Liu, H. Feature selection for classification. *Intelligent Data Analysis* (1997) doi:10.3233/IDA-1997-1302.

38. Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. in *Data Classification: Algorithms and Applications* (2014). doi:10.1201/b17320.

39. Jovic, A., Brkic, K. & Bogunovic, N. A review of feature selection methods with applications. in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 1200–1205 (IEEE, 2015). doi:10.1109/MIPRO.2015.7160458.

42. Longadge, M. R., Snehlata, M., Dongre, S. & Latesh Malik, D. *Class Imbalance Problem in Data Mining: Review*. *International Journal of Computer Science and Network* vol. 2 www.ijcsn.org (2013).

43. Wong, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* **48**, 2839–2846 (2015).

44. Tharwat, A. Classification assessment methods. *Applied Computing and Informatics* **17**, 168–192 (2021).

45. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**, 197–227 (2016).

46. Adnan, N., Ahmad, M. H. & Adnan, R. *A Comparative Study On Some Methods For Handling Multicollinearity Problems*. *MATEMATIKA* vol. 22 (2006).

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

47. Beretta, L. & Santaniello, A. Nearest neighbour imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making* **16**, 74 (2016).

48. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research* **20**, 40–49 (2011).

49. Poulos, J. & Valle, R. *MISSING DATA IMPUTATION FOR SUPERVISED LEARNING †*. (2018).

50. Watts, D. *et al*. Predicting offences among individuals with psychiatric disorders - A machine learning approach. *Journal of Psychiatric Research* **138**, 146–154 (2021).

Chapter 3 - Predicting criminal and violent outcomes in psychiatry: a meta-analysis of diagnostic accuracy

Authors: Devon Watts, MSc[1,2]; Taiane de Azevedo Cardoso, MSc, PhD[1], Diego Librenza-Garcia[1,3] MD, PhD; Pedro Ballester MSc[1,2]; Ives Cavalcante Passos MD, PhD[4,6]; Felix H. P. Kessler PhD[5]; Jim Reilly, PhD[8]; Gary Chaimowitz, MB, ChB, FRCP[1,7]; Flavio Kapczinski MSc, MD, PhD, FRCPC[1,2,6]

1. Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada.
2. Neuroscience Graduate Program, McMaster University, Hamilton, ON, Canada.
3. Post-Graduation Program in Psychiatry and Behavioural Sciences, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil.
4. Laboratory of Molecular Psychiatry, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, RS, Brazil.
5. Center for Drug and Alcohol Research, HCPA, Porto Alegre, RS, Brazil.
6. Instituto Nacional de Ciência e Tecnologia Translacional em Medicina (INCT-TM), Porto Alegre, RS, Brazil.
7. Forensic Psychiatry Program, St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada.
8. Department of Electrical and Computer Engineering, McMaster University, Canada.
*Corresponding author:
Flavio Kapczinski, MSc, MD, PhD, FRCP(C)
Director, Center for Clinical Neuroscience
Professor, Department of Psychiatry
Psychiatry & Behavioural Neurosciences, McMaster University
100 West 5th Street, Hamilton, Ontario, L9C 0E3, Canada
Phone: 905-522-1155 ext. 35420
Email: kapczinf@mcmaster.ca
ORCID: 0000-0001-8738-856X

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

Email for all authors:
Devon Watts: wattsd@mcmaster.ca
Taiane de Azevedo Cardoso: deazevet@mcmaster.ca
Diego Librenza-Garcia: librenzagarcia@gmail.com
Pedro Ballester: ballestp@mcmaster.ca
Ives Cavalcante Passos: ivescp1@gmail.com
Felix H. P. Kessler: fkessler@hcpa.edu.br
Jim Reilly: reillyj@mcmaster.ca
Gary Chaimowitz: chaimow@mcmaster.ca
Word count (including citations): 6312

This chapter in its entirety has been ***published*** in **Translational Psychiatry**. The final accepted manuscript version of this article is presented within this thesis.

Watts, D., de Azevedo Cardoso, T., Librenza-Garcia, D. *et al.* Predicting criminal and violent outcomes in psychiatry: a meta-analysis of diagnostic accuracy. *Transl Psychiatry* **12**, 470 (2022). https://doi-org.libaccess.lib.mcmaster.ca/10.1038/s41398-022-02214-3

Abstract

Although reducing criminal outcomes in individuals with mental illness have long been a priority for governments worldwide, there is still a lack of objective and highly accurate tools that can predict these events at an individual level. Predictive machine learning models may provide a unique opportunity to identify those at highest risk of criminal activity and facilitate personalized rehabilitation strategies. Therefore, this systematic review and meta-analysis aims to describe the diagnostic accuracy of studies using machine learning techniques to predict criminal and violent outcomes in psychiatry.

We performed meta-analyses using the mada, meta, and dmetatools packages in R to predict criminal and violent outcomes in psychiatric patients (n=2428) (Registration Number: CRD42019127169) by searching PubMed, Scopus, and Web of Science for articles published in any language up to April 2022.

Twenty studies were included in the systematic review. Overall, studies used single-nucleotide polymorphisms, text analysis, psychometric scales, hospital records, and resting-state regional cerebral blood flow to build predictive models. Of the studies described in the systematic review, nine were included in the present meta-analysis. The area under the curve (AUC) for predicting violent and criminal outcomes in psychiatry was 0.816 (95% Confidence Interval (CI): 70.57-88.15), with a partial AUC of 0.773, and average sensitivity of 73.33% (95% CI: 64.09-79.63), and average specificity of 72.90% (95% CI: 63.98-79.66), respectively. Furthermore, the pooled accuracy across models was 71.45% (95% CI: 60.88-83.86), with a tau squared (τ^2) of 0.0424 (95% CI: 0.0184-0.1553).

Based on available evidence, we suggest that prospective models include evidence-based risk factors identified in prior actuarial models. Moreover, there is a need for a greater emphasis on

identifying biological features and incorporating novel variables which have not been explored in prior literature. Furthermore, available models remain preliminary, and prospective validation with independent datasets, and across cultures, will be required prior to clinical implementation. Nonetheless, predictive machine learning models hold promise in providing clinicians and researchers with actionable tools to improve how we prevent, detect, or intervene in relevant crime and violent-related outcomes in psychiatry.

Keywords

machine learning; precision psychiatry; artificial intelligence; forensic psychiatry; psychotic disorders; computational psychiatry; criminality; diagnostic accuracy

3.1 Introduction

Available evidence suggests that one in eight men, and one in sixteen women will subsequently commit a serious criminal offense after release from a psychiatric facility [1]. This phenomenon is not isolated to specific geographical or generational effects, considering that in a systematic review comprising 33,588 individuals from 24 countries and 109 datasets, high rates of mental illness in prisoners were found in both high- and low-income countries over the timespan of four decades [2].

Additionally, results from a large Swedish registry study comprising 98,082 individuals with a history of hospitalization suggests that one in every twenty violent crimes is committed by someone with severe mental illness [3]. Given the high prevalence of criminal acts committed across cultures in individuals with severe mental illness, there has been a concerted effort to identify predictors of prospective criminal risk following discharge from psychiatric facilities.

In response to this, actuarial assessments became increasingly widespread, which use statistical algorithms to identify prospective patient risk, usually at the group level [4]. However, there is little evidence that actuarial risk estimates can accurately determine whether a specific patient will reoffend or commit subsequent acts of violence [5]. This is largely because most risk estimates have been developed statistically to assess group-based risk and perform poorly when making individualized predictions [5]. Altogether, this illustrates the limitations of current methods and the importance of a more precise, effective, and personalized approach to risk assessment in forensic settings. Given the ethical, psychiatric, and legal ramifications of inappropriately mischaracterizing the prospective risk of any given patient, and the resulting consequences to the individual, their families, and broader society, there is a growing interest in the use of artificial intelligence and predictive analytics to facilitate clinical decision making at an individual level [11]. This can potentially pave the way for tailor-made tools for the diagnosis, assessment, and treatment of patients [6,7]. While predictive machine learning models have already shown promise in other fields of medicine [8,9], there is a growing effort towards predicting criminal outcomes in psychiatric patients at an individual level. Incorporating such models into routine clinical care presents with the potential to facilitate personalized and targeted rehabilitation strategies to decrease prospective criminal outcomes. To the best of our knowledge, there are no systematic reviews describing the diagnostic accuracy of machine learning models in predicting criminal and violent outcomes in psychiatry. Therefore, this systematic review and meta-analysis aims to assess the diagnostic accuracy of studies using machine learning techniques to predict criminal outcomes in psychiatry.

2.2 Methods

This study has been registered on PROSPERO with the registration number PROSPERO CRD42019127169.

3.2.1. Search strategy

We searched three electronic databases (PubMed, Scopus, and Web of Science) for articles published up until April 2022. To identify relevant studies, the following structure for the search terms was used: (Artificial Intelligence OR Supervised Machine Learning AND crime-related outcomes in psychiatry). The complete search filter is available in the supplementary material. We also screened references from included articles to search for potentially missed articles.

3.2.2. Eligibility criteria

This systematic review was performed according to the PRISMA statement [10]. We selected original articles that used supervised machine learning models to predict crime-related outcomes in mental illness. We excluded review articles and studies using unsupervised learning, since methods such as clustering are not outcome oriented. Furthermore, studies that predicted crime or violent-related outcomes in individuals without psychiatric disorders were excluded, although further information regarding these studies can be found in Supplementary Table 2.

3.2.4. Data collection and extraction

Potential articles were independently screened in a blinded standardized manner for title and abstract contents by two researchers (DW and DLG). Following this, the full texts of screened articles were obtained and evaluated according to the inclusion and exclusion criteria. A third author (PB) provided a final decision in cases of disagreement. Criminal outcomes were

operationalized as rearrest, reconviction of crimes, or prediction of the type of crime committed. Violent outcomes involved recorded violent incidents during inpatient stay or following hospital discharge.

3.2.5. Quality assessment

We created a machine learning quality assessment table based on experts' opinion to evaluate the reproducibility and reliability of the included studies. Our assessment provides a quick way to evaluate published papers and can also serve as a checklist for future studies. Briefly, the instrument comprises nine methodological considerations, including representativeness of the sample, confounding variables, outcome assessment, algorithm selection, feature selection, class imbalance (where applicable), missing data, performance/accuracy, and testing/validation. The instrument can be found in Supplementary Table S1, and further details can also be found in the Supplementary Material.

3.3. Statistical analysis

A bivariate meta-analysis was performed for crime-related and violent outcomes using the mada [11] meta [12], and dmetatools packages in R [11]. Since we anticipated considerable between-study heterogeneity, a random-effects model was used to pool effect size. Additionally, an adjusted profile restricted maximum likelihood estimator was used to calculate the heterogeneity variance tau square (τ^2). This metric was selected since the heterogeneity statistic $\boldsymbol{I^2}$ can be biased in meta-analyses with small sample sizes [13]. Using the retisma function in 'mada' [11], a linear mixed model with random effects was selected to produce summary estimates of sensitivity and specificity, as well as calculate AUC and partial AUC summary receiver operating characteristic (ROC) curves, as described elsewhere [14]. 95% confidence intervals for summary AUC were

generated using 2000 iterations of parametric bootstrapping with the 'dmetatools' package in R. Additionally, using the metamean function in 'meta' [12], mean accuracy across models was pooled alongside standard error of model accuracy, as detailed in Supplementary Table S3. As we anticipated considerable between-study heterogeneity, a random-effects model was selected to pool effect sizes. The restricted maximum likelihood estimator [15] was selected to calculate the heterogeneity variance τ2. Knapp-Hartung adjustments [16] were also used to calculate the confidence interval around the pooled effect. Additionally, we pooled the diagnostic odds ratio, and the positive negative and likelihood ratios within a random effects model with a DerSimonian-Laird estimator [17].

Four studies were excluded from the meta-analysis, as the authors did not report the sensitivity and specificity of their models. Criminal outcomes were operationalized as rearrest, reconviction of crimes, or prediction of the type of crime committed. Violent outcomes involved recorded violent incidents during inpatient stay or following hospital discharge.

3.4. Results

We found 12420 potential titles/abstracts and included 20 studies which met inclusion criteria. A list of the included studies and their most relevant characteristics and findings are described in Table 1, while Table 2 details the diagnostic accuracies, odds ratios, and likelihood ratios of studies contained within the meta-analysis. Additionally, a schematic of the meta-analytic diagnostic accuracy of predicting criminal recidivism and physical violence are detailed Figure 1. Furthermore, a machine learning quality assessment, additional figures related to model performance, and a table comprising twenty-one studies assessing criminal outcomes in non-psychiatric individuals can be found in the supplementary material. Additional information about

machine learning algorithms [18] including methodological considerations, common problems, and limitations, can be found elsewhere [19].

Of the studies included in the systematic review, six assessed predictors of criminal recidivism [20–25], two assessed predictors of the type of criminal offence [26,27], three assessed predictors of physical violence during inpatient stay [28–30], and six assessed predictors of violent offending and aggression following discharge [24,31–38]. All studies, apart from two [21,30], used clinical input features, including socio-demographic information, questionnaires, and psychometric measures to derive predictions.

3.4.1. Studies assessing criminal outcomes

Eight studies used machine learning models to predict criminal outcomes in patients with psychiatric disorders [20–27]. Delfin and colleagues conducted the first 10-year follow-up of a cohort of forensic psychiatry patients, including 44 individuals, who underwent a single-photon emission CT scan. This data, alongside eight evidence-based clinical risk factors, were used in a random forest model to predict criminal recidivism, resulting in an accuracy of 82% and an AUC of 0.81. Of note, when only clinical risk factors were used alone, model performance degraded, with an accuracy of 64% and AUC of 0.69, emphasizing the importance of combining clinical and biological features to predict criminal recidivism. The top features reflecting neuronal activity included the right and left parietal lobe, left temporal lobe, and right cerebellum [21].

Kirchebner and colleagues used 653 clinical features to predict recidivism in 344 individuals with schizophrenia. Patients who had a criminal record prior to their current offence were considered as recidivists. Following imputation, the best performance was observed using Boosted Trees, with an accuracy of 67.6%. Without imputation, a Naive Bayes classifier

achieved an accuracy of 79.4%. Important variables included amisulpride prescription prior to offence, recent stressors, recent legal complaints, and number of prior offences [24].

Sonnweber et al. developed a model to differentiate between violent and non-violent offenders in patients with schizophrenia. The best performance was observed using a gradient boosting machine, resulting in a balanced accuracy (operationalized as the average of sensitivity and specificity, as defined elsewhere [39]) of 67%. The most important variables included time spent in hospitalization, age at diagnosis, daily olanzapine at discharge, PANSS score at discharge, and social isolation in adulthood [26].

Furthermore, Watts and colleagues developed a machine learning model to predict the type of criminal offence committed in a large transdiagnostic sample of 1240 psychiatric patients. Using multiclass classification, they showed that sexual crimes could be discriminated from violent and nonviolent crimes at an individual level with an accuracy of 71.22%. Moreover, following recursive feature elimination, a reduced model with 36 variables resulted in an accuracy of 71.58%. The most important features for the model included previous absolute discharge, previous sexual convictions, cluster A personality disorder, and female gender [27]. Other studies predicted rearrest after release from jail [20,22], reconviction for a violent crime [23], and risk of general criminal recidivism [25]. A summary of these findings can be found in Table 1 and Supplementary Table S2.

3.4.2. Studies assessing violent outcomes

Twelve studies used machine learning techniques to predict violent outcomes in patients with psychiatric disorders [28–38,40]. Linaker and colleagues predicted violent incidents in psychiatric patients using behavioral symptoms from health records from 24 hours prior. Overall, 48 acts of

violence were recorded from 32 patients, and following feature selection using correlation coefficients, six variables were used as predictors in a logistic regression model. The authors reported a sensitivity of 81.3% and specificity of 100%, however it was unclear how class imbalance was addressed, since only 34.7% of patients committed an act of violence during the study [32].

Kirchebner and colleagues used a series of known stressors to predict violent offending in 370 patients with schizophrenia. The overarching goal was to determine whether accumulated stressors precipitated violent outcomes in patients. Using boosted classification trees, they reported an accuracy of 76.4%. However, no external validation or testing set was used, instead, performance was assessed using 5-fold CV [40].

Furthermore, Menger et al. used text analysis from doctor and nurse notes to predict violent incidents in psychiatric inpatients. Four feature extraction methods were used, comprising binary bag of words, term frequency-inverse document frequency (tf-idf) bag of words, document embeddings, and word embeddings, as described elsewhere. An AUC of 0.788 was observed using document embeddings with recurrent neural networks. The worst performances occurred with the Naive Bayes algorithm, which is the most classical and widely used algorithm for text classification [28].

Monahan and colleagues classified patients according to high and low risk of violence following discharge from psychiatric facilities. Decision trees were used in a binary classification task, and features were selected using a stepwise model, where the threshold of statistical significance between the feature and outcome were set at $P<0.05$. The model correctly identified 72.6% of the sample as either low or high risk. Important variables included seriousness of prior arrests, motor impulsiveness, paternal drug use, and recurrent violent fantasies. It is important to mention that

27.4% of the total sample remained unclassified, meaning it could find no combination of risk factors to classify patients into high or low-risk groups [33].

Additionally, Suchting and colleagues used saliva FK506 binding protein 5 (FKBP5) polymorphisms alongside demographic and psychometric variables to predict state aggression, which resulted in an R^2 of 0.66 [30]. Other studies identified predictors of violent risk following discharge [37,38] and aggression in patients [29,31,34–36], which are further described in Table 1.

3.4.3. Meta-analysis of diagnostic accuracy

A forest plot detailing model performance can be observed in Figures 1 and 2, while Table 2 details the diagnostic accuracies, odds ratios, and likelihood ratios across studies. Additional details related to the standard error of model accuracy, 95% CI, and the true/false positives and negatives, can be found in Supplementary Table S3. Nine studies were pooled, comprising 2,428 patients (the same dataset of 370 patients was used across two studies [26,40]).

Additionally, nine studies which did not report the sensitivity and specificity of models [20,22,23,28,29,31,33–35], and one regression-based model [30] were excluded from the meta-analysis. Overall, the pooled accuracy across models was 71.45% (95% CI: 60.88-83.85), with a sensitivity ranging from 54.4%-87.3% (average: 73.33%, 95% CI: 64.09-79.63) and specificity ranging from 60.5-96.6% (average: 72.90%, 95% CI: 63.98-79.66). The heterogeneity statistic τ^2 for pooled model accuracy was 0.0424 (95% CI: 0.0184-0.1553). A plot of the false positive rate against sensitivity for all studies can be found in Supplementary Figure S1.

The diagnostic odds ratio (DOR) across studies was 9.75 (95% CI: 4.035-22.72; τ^2=1.505) as detailed in Table 2. Similarly, the positive likelihood ratio (posLR) was 3.083 (95% CI: 1.954-4.866, with a τ^2 of 0.437 (95% CI: 0.000-0.897), and the negative likelihood ratio (negLR) was 0.342 (95% CI: 0.201-0.583), with a τ^2 of 0.566 (95% CI: 0.000-3.476), respectively.

Additionally, the log DOR across studies was 2.466 (95% CI: 1.534-3.397). The average prevalence of the positive class (presence of criminal and violent outcomes) was 43.435% of the sample across studies. Furthermore, the AUC across studies was 0.816 (95% CI: 0.745-0.875) in predicting criminal and violent outcomes, with a partial AUC of 0.773. Spearman's rho indicated a weak association (rho=0.150, 95% CI: -0.571-0.740) with a large confidence interval between the sensitivities and false positive rates of included studies.

3.5. Discussion

To the best of our knowledge, this is the first systematic review and meta-analysis comprising studies using supervised machine learning techniques to predict criminal or violent outcomes in individuals with psychiatric disorders. Throughout our review, we have identified recurrent features and algorithms used, as well as current methodological challenges. In this section, we detail key aspects of these models, showcasing their limitations as well as our perspectives on best practices for developing machine learning models with clinical utility. Further details regarding common methodological issues in machine learning models can be observed in the supplementary material.

3.5.1. Model interpretability, model performance, and confidence intervals

More recent machine learning algorithms that use regularization parameters to account for common issues such as multicollinearity, tended to show higher performance accuracy in predicting outcomes. However, model complexity carries the trade-off of greater difficulty in model interpretability and explainability [41].

Recently, new local explanation methods have been developed, including SHapley Additive exPlanations (SHAP), to explain variable contributions at the individual level [42]. Adaptations of this, such as TreeExplainer, leverage the internal structure of tree-based models to efficiently compute local explanations using Shapley values [43]. Moreover, SHAP dependence plots can be used to showcase the effect that a single feature has on predictions made by the model [43]. In two studies included in the current review, feature importance metrics were not reported [28,35]. It is argued that future studies may benefit from an increased focus on model interpretability, which may aid in the generalizability and replicability of such work.

Furthermore, it is important to highlight that model performance can be over-optimistic when assessed using internal cross-validation alone, in the absence of separate training and testing sets. Of the twenty studies contained in the present review, only seven (35%) incorporated training and testing sets in model development. In the majority of studies [25,28–31,33–36,38] (76.9%) that evaluated model performance using internal cross-validation alone, sample sizes were also well over 100 patients. As mentioned elsewhere, several other fields use cross-validation to tune regularization parameters in model development, rather than taking performance estimates at face value [44]. Similarly, it is important to mention that uncertainty estimates should be considered when evaluating model performance and its potential clinical utility. Of nine studies comprising the meta-analysis, only four (44.4%) [21,26,27,37] reported accuracy estimates using a method such as 95% confidence intervals.

3.5.2. Model Performance and Clinical Predictors

Overall, eighteen models assessed clinical predictors of criminal and violent outcomes [20,22,32–38,40,23–29,31]. In criminal prediction models, accuracy was generally high, ranging from 67.83-82%.

With respect to criminal behavior, common predictors across models included age at first crime, substance use disorder, cluster B personality disorder, prior criminality, a high number of stressors, and childhood trauma. Future work may benefit from comprising a standardized evidence-based risk battery for use in prospective models.

Furthermore, models predicting violent behaviour were more variable, ranging from 58.25-92.1%, with five of twenty studies (25%) [22,23,28,35] comprising the systematic review only reporting AUC. As such, several were excluded from the meta-analysis. Nonetheless, important clinical features included confusion, irritability, threats, recently attacking objects, child abuse, physical neglect, and callous affect. Important search terms included aggressive, offered, angry, door, walk, arrest, offer emergency medication, and walked.

With respect to the meta-analysis comprising nine studies (n=2,428 patients), the pooled accuracy was 71.45% (95% CI: 60.88-83.86) in predicting criminal and violent outcomes. Moreover, as detailed in Table 2, the DOR was 9.757 (95% CI: 4.035-22.72; τ^2= 1.505) and log DOR was 2.466 (95% I: 1.534-3.397). As discussed elsewhere, the DOR is a measure of the effectiveness of a diagnostic test that is independent of prevalence [45]. A DOR of 9.757 represents a high ratio of the odds of the test being positive if the individual will commit prospective criminal and violent outcomes relative to the odds of the test being positive if the individual will not prospectively commit criminal and violent outcomes. However, a large upper and lower bound of the 95% CI was observed, and the log DOR suggests a more conservative test effectiveness. Similarly, the posLR was 3.083 (95% CI: 1.954-4.866), suggesting a small increase in the likelihood of committing violent and criminal outcomes in patients with a positive test. In addition, the negative likelihood ratio was 0.342 (95% CI: 0.201-0.583), suggesting a 20-

25% decrease in the odds of committing violent and criminal outcomes in patients with a negative test result.

3.5.3. Model Performance and Biological Predictors

Furthermore, two models [21,30] assessed biological predictors pertaining to saliva SNPs and resting-state regional cerebral blood flow. Although they contained small sample sizes and lacked external validation, both showed promising performance, corresponding to an R2 of 0.66, and accuracy of 82%, respectively. Important features included KBP5_14 (rs1460780), FKBP5_92 (rs9296158); and FKBP5_94 (rs9470080), right and left parietal lobe rCBF, left temporal lobe rCBF, and right cerebellum. Subsequent studies may benefit from replicating these findings and incorporating additional biological and physiological variables.

3.5.4. Limitations

Currently, the field of predicting crime and violent related outcomes using machine learning techniques remain in its infancy. As such, there is a lack of studies validating model performance using independent cohorts. Furthermore, it is important to note that model accuracy should be considered alongside several other factors, such as the input features used, the preprocessing pipeline, feature selection method, model optimization strategy, and the validation procedure. Furthermore, data-driven approaches to feature selection can be useful in many cases, since it does not require knowledge derived from pre-existing literature to manually select important variables [46–48]. Of note, the absence of a formalized feature selection strategy was observed across a subset of studies.

There are several available feature selection methods, with varying degrees of appropriateness depending on the application, as described elsewhere [47]. Furthermore, feature selection can be useful to improve the generalizability of models when applied to independent datasets [49]. Considering that predictive models applied to forensic healthcare can have significant legal repercussions - such as incorrectly identifying individuals as not criminally responsible when in fact they are, or the inability to detect malingering - it is paramount that we use the most optimal methods available for these purposes.

Additionally, only two studies developed separate models to assess potential differences in performance between men and women using the same variables, as described in the supplementary material. Rosselini et al. reported an AUC of 0.74 for men and an AUC of 0.82 for women in predicting violent crime [50]. Additionally, the same authors also investigated predictors of major violent crime and reported an AUC of 0.81 for both models in men, and an AUC of 0.80-0.82 for both models in women. Based on these studies, it is still unclear whether biological sex or gender play a key role in deciding which features should be included within a predictive machine learning model.

3.5.5. Future directions

Moving forward, a further refinement of predictive models in forensic risk prediction is required. Potentially, this may be facilitated by using a wider framework when selecting the input data in our models. Considering that our model performance is directly dependent on the available input data, an exploratory data-driven approach may be warranted in predictive models.

Most machine learning studies in forensic psychiatry thus far focus purely on clinical and administrative data, given the widespread availability of such data. However, other modalities, such as neuroimaging (MRI, fMRI, DTI), electrophysiology (EEG, MEG, ERG) various sensors

(actigraphy, heart rate variability), and genomic features (whole genome sequencing, whole exome sequencing, and RNA sequencing) may prove to facilitate model performance, when used in conjunction with clinical data. Moreover, longitudinal studies with larger multicentric samples and adequate external validation are needed to translate proof-of-concept predictive models into applications to be used in clinical and legal settings. We hypothesize that such models may facilitate a more personalized approach to patient evaluation and risk management, provide greater precision in deriving a tailored treatment plan, and aid clinicians and the legal system in the decision-making process as it pertains to mentally disordered offenders. Ultimately, they may become critical tools to assist in prison sentencing, to determine fitness to stand trial, and to optimize the progress of individuals in the forensic system towards rehabilitation.

Author's Contributions

Devon Patrick Watts, Taiane de Azevedo Cardoso, Diego Librenza-Garcia, and Pedro Ballester participated in the literature search, writing, and in the approval of the final manuscript. Ives Cavalcante Passos, Felix H. P. Kessler, Jim Reilly, Flavio Kapczinski, and Gary Chaimowitz participated in the writing and in the approval of the final manuscript.

Conflict of interest statements

Devon Watts reports a PhD fellowship from the Canadian Institute of Health Research (CIHR), outside the submitted work. Taiane de Azevedo Cardoso reports a postdoctoral fellowship from the Canadian Institute of Health Research (CIHR), outside the submitted work. Ives Cavalcante Passos reports consulting fees from Torrent/Omnifarma, and previous funding from INCT - CNPqand CAPES. Flávio Kapczinski reports personal fees from Daiichi sankyo, and Janssen-Cilag; grants from Stanley Medical Research Institute 07TGF/1148, grants from INCT - CNPq 465458/2014-9, and from the Canadian Foundation for Innovation - CFI, outside the submitted

work. Diego Librenza-Garcia, Pedro Ballester, Felix Kessler, Jim Reilly, and Gary Chaimowitz report no biomedical financial interests or potential conflicts of interest.

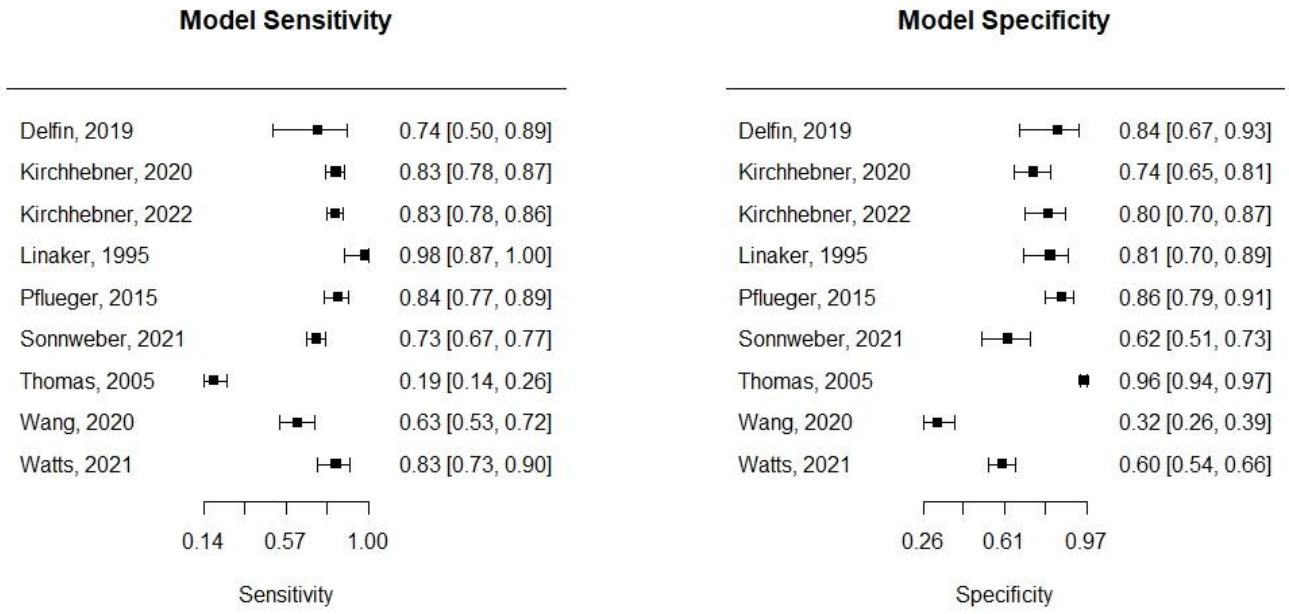


Figure 1: *Paired Forest plot of model accuracy for criminal and violent outcomes in psychiatry*
A linear mixed model with random effects was selected to produce summary estimates of sensitivity and specificity using the retisma function in mada. The average sensitivity across studies was 73.33% (95% I: 64.09-79.63) and average specificity was 72.90% (95% CI: 60.50-96.6). As such, the balanced accuracy across models (sensitivity + specificity / 2) is 73.11%.

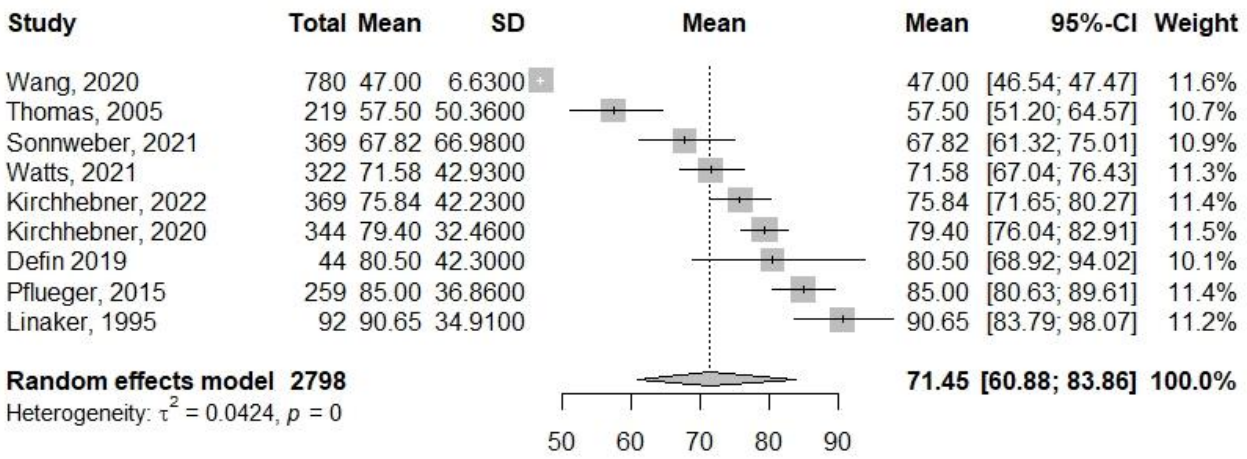


Figure 2: Pooled Effects of Model Accuracy

Pooled accuracy of criminal and violent models in psychiatry across 2428 patients (two studies used the same sample n=370) within a random-effects model using a restricted maximum likelihood estimator to calculate the heterogeneity variance τ2. Reported mean accuracy across models was used, in conjunction with standard deviation, calculated by multiplying the standard error by the square root of the sample size (SD = SE×√n). Knapp-Hartung adjustments were used to calculate the confidence interval around the pooled effect. The average accuracy across models was 71.45% (95% CI: 60.88-83.86), with a heterogeneity variance τ2 of 0.0424.

First author, year	Data utilized	Outcome	Sample size and diagnosis[1]	Validation	Machine learning model	Accuracy	Other measures
CRIMINAL OUTCOMES							
Cohen 1988	Clinical and administrative data	Subsequent arrest	127 male patients found not guilty by reason of insanity	N/A	Stepwise discriminant analysis	76%	N/A
Delfin 2019	resting-state regional cerebral blood flow (rCBF) and clinical risk factors	Criminal recidivism	44 forensic psychiatry patients	Out-of-bag (OOB) error	RF	Accuracy: 82%	AUC: 0.81 Sensitivity: 75% Specificity: 86% PPV = 0.73 NPV = 0.86 Note: the dataset was not split into training and testing sets, and OOB error was used as a resampling procedure
Falconer, 2014	Age, past arrests, mental health diagnosis, enrollment to the JDP as well as utilization of outpatient group services, medical services, and case management	Rearrest	2100 adult offenders with records in US mental health services and the criminal justice system	Training (80%) and testing (20%) sets	Elastic Net regularized logistic regression	N/A	AUC (test set) 0.67 0.60 (simplified model)
Grann, 2007	10 risk factors of the Historical subscale of the HCR-20	Reconviction for a violent crime	404 violent offenders with a mental disorder followed up to eight years	Holdout validation with training/testing (2:1) (ANN)	BLR MLR ANN	N/A	AUCs BLR: 0.66-0.77 MLR: 0.63-0.73 ANN: 0.51-0.73

Kirchebner, 2020	Sociodemographic, clinical, behavioral, and symptom variables	Criminal recidivism	344 offenders with schizophrenia	Training (70%) and testing (30%) sets	Boosted Trees Naive Bayes	67.6-79.4% *Best performance using Naive Bayes*	*Best model* Naive Bayes with imputation AUC: 0.83 Sensitivity: 83% Specificity: 74% PPV: 84% NPV: 73%
Pflueger, 2015	Demographic variables and clinical scales (Basel Catalog for Risk Assessment, Historical Clinical Risk Assessment, and the Psychopathy Checklist-screening version)	Risk of general criminal recidivism of offenders with mental illness	259 individuals subjected by court orders to forensic psychiatric evaluation for mental and behavioral disorders using the ICD-10	4-fold cross-validation	RF	Best model had an overall 85% accuracy and accounted for 91% of all observed re-offenses.	Best model had a sensitivity of 84% and specificity of 86%.
Sonnweber, 2021	Clinical, developmental and social factors	Discriminating between violent and nonviolent offending	370 forensic offenders with schizophrenia	Training (70%) and testing (30%) sets	LR RF GBM KNN SVM Naive Bayes	Best model had a balanced accuracy of 67.82%	Sensitivity: 72.73% Specificity: 62.92% PPV: 65.98 NPV: 70.00 AUC: 0.764
Watts, 2021	Sociodemographic, clinical, behavioral, and symptom variables	Type of criminal offence (violent, sexual, nonviolent)	1240 transdiagnostic patients	Training (70%) and testing (30%) sets	RF Elastic Net SVM	Violent vs Sexual Offences: 65.27-80.31% Nonviolent vs Sexual Offences: 49.56-77.62% Sexual Offences vs Violent and Nonviolent: 59.82-71.58%	Best models: Violent vs Sexual Offences: Sensitivity: 76.74% Specificity: 83.87%

							PPV: 97.06 NPV: 34.21 Nonviolent vs Sexual Offences: Sensitivity: 74.60% Specificity: 80.65% PPV: 80.65% NPV: 60.98% Sexual vs Nonviolent and Violent Offences: Sensitivity: 83.15% Specificity: 60.00% PPV: 95.08 NPV: 27.69
VIOLENT OUTCOMES							
Kirchebner 2022	Clinical variables pertaining to childhood, adolescence, adulthood and psychiatric stressors	Violent offending in schizophrenia	370 offenders with schizophrenia	5-fold cross-validation; no external validation used.	Boosted Classification Trees	76.4%	Sensitivity: 80.49 Specificity: 71.19 PPV: 66 NPV: 84 AUC: 0.83
Le, 2018	Text analysis from electronic mental health records	Forensic risk assessment ratings as a proxy of violence to others	Four NLP dictionary word lists - 6865 mental health symptom words from Unified Medical Language	10-fold stratified cross-validation; no external validation used.	Bagging J48 JRip	SVM and LMT were the most accurate algorithms (accuracy of 69-77%) with all three dictionaries.	N/A

			System, 455 DSM-IV diagnoses from UMLS repository, 6790 English positive and negative sentiment words, and 1837 high-frequency words from the Corpus Contemporary American English (COCA). *Exact number of patients not reported*		LMT LR Linear Regression SVM		
Linaker, 1995	55 items describing symptoms or behaviors reported or believed to be positively or negatively related to violent behaviors, obtained through screening of the medical records in the 24 hours prior to the outcome	Physical violence towards others, assessed by the screening of medical records	94 patients admitted to a maximum-security psychiatric unit	Holdout validation, with training (46.1%) and testing (53.9%) sets	LR	92.1%	Specificity 100% Sensitivity 81.3%
Menger, 2018	25.942 doctor and nurse text notes at the start of admission (predictors) and violence incident reports (outcome)	Violent incidents in an inpatient unit occurring within the first 30 days of admission	2521 psychiatric admissions from 6 inpatients units	5-fold cross-validation. no external validation	RNN CNN NN NB SVM DT	N/A	AUCs ranged from 0.654 (word embeddings with RNN) to 0.788 (documents embedding with RNN)
Menger, 2019	Electronic health records	Inpatient violent risk	2209 psychiatric patients	Training (53.5%) and testing (46.5%) samples	SVM (radial kernel)	Testing / Validation (Sensitivity/Specificity) Site 1: 92.5% / 24.8%	AUC Site 1: 0.722 (0.690-0.753 95% CI) Site 2: 0.643 (0.610-0.675 95%

						Site 2: 92.9% / 13.4%	CI)
Monahan, 2000	Clinical data obtained from interview, records, and questionnaires	Violent incidents after 20 weeks of hospital discharge	939 psychiatric inpatients	Bootstrapping	ICT	N/A	72.6% of the sample classified as low or high risk based on the prevalence of incident events based on a cut-off stipulated by the authors
Steadman, 2000	Clinical and demographic risk factors collected through the MacArthur Violence Risk Assessment Study	Predictors of violence risk	939 psychiatric patients assessed during the first 20 weeks following hospital discharge	Bootstrapping (1000 random samples with replacement drawn from original sample of 939).	LR CTA ICT	N/A	LR: 0.81 AUC CTA: 0.79 AUC ICT: 0.82 AUC Did not report sensitivity, specificity, PPV or NPV.
Suchting, 2018a	Demographic variables, psychometric variables, and saliva samples for genetic testing of FKBP5 SNPs (FKBP5_13 (rs1360780); FKBP5_92 (rs9296158); and FKBP5_94 (rs9470080).	Predictors of State Aggression in individuals with previous trauma	48 participants selected irrespective of DSM diagnostic or psychometrically established clinical cut-offs for trauma exposure.	10-fold cross-validation; no external validation used.	Component-wise gradient boosting; backward elimination used for feature selection.	N/A	8-factor model R^2 =0.66 Did not report AUC, accuracy, sensitivity, specificity, PPV or NPV.
Suchting, 2018b	Extracting variables using retrospective electronic health records	Predictors of aggression in inpatients	29,841 patient records from the Harris County Psychiatric Center	10-fold cross-validation; no external validation used	Four different algorithms: GLM RF GBM DNN	N/A	GLM: 0.7801 AUC RF: 0.7420 AUC GBM: 0.7765 AUC DNN: 0.7137 AUC
Thomas, 2005	Data from a large randomized controlled trial in	Predictors of violence among patients with	780 patients with psychosis, 158 of	10-fold cross-validation; no	Full logistic regression (14	57.5%	*Best Performance*

	4 inner-city mental health services in the United Kingdom (clinical/demographic variables)	psychosis	which were violent during the 2-year follow-up period	external validation used	variables) Forward stepwise logistic regression (6 variables) Full CART (123 nodes) Pruned CART (22 nodes) Pruned CART (22 nodes: violent cases given, 5 x weight)		Full logistic regression Sensitivity - 19% Specificity - 96% PPV - 49% NPV - 79% Percent correctly classified - 77%
Tzeng, 2004	Patient insight ratings, medication compliance, and demographic characteristics Schedule for Assessment of Insight in Psychosis (SIP) Violence and Suicide Assessment Scale (VASA)	Presence or absence of violent behavior towards people or things (1 year later)	63 outpatients with schizophrenia, according to the DSM-IV, who were in remission or had minimal psychosis symptoms	3-fold cross-validation; no external validation used	SVM	76.2%	An LR model was used as a point of comparison, however, no resampling measures were used (model developed using the entire sample)
Wang, 2020	Identified 28 variables previously identified with violence or schizophrenia (Structured interviews, self-report questionnaires, medical history, and demographic information)	Violent vs Nonviolent (Ranging from absence of physical violence to assault causing bodily harm according to the Modified Overt Aggression Scale)	275 patients with schizophrenia spectrum disorder, according to the DSM-IV	5-fold cross validation; no external validation used	LR LASSO Elastic Net RF GBRT	57-62% *Best performance using RF*	*Best performance* *Random Forest* AUC: 0.63 (± 0.004) Sensitivity: 63% (± 0.005)

					SVM *radial kernel*		Specificity: 32% (± 0.008) PPV: 62% (± 0.008) NPV: 54% (± 0.003)

Table 1 – Predicting criminal and violent outcomes in psychiatry.

A summary of input data, sample characteristics, validation methods, and machine learning models across studies.

Abbreviations:

ANN, Artificial neural networks; *AUC*, Area under the curve; *CART*, Classification and regression trees; *CNN*, Convolutional neural networks; *CTA*, Classification Tree Analysis; *DNN*, Deep neural networks; *DSM IV-R*, Diagnostics and Statistical Manual, Version IV, Revised; *DT*, Decision tree; *EN*, elastic net; *GBRT*, Gradient Boosted Regression Trees; *HCR-20*, Historical, clinical, risk management-20; *ICT*, Iterative classification tree; *LASSO*, Least Absolute Shrinkage and Selection Operator; *LR*, Logistic regression; *NB*, Naive Bayes; *NN*, Neural network; *NPV*, Negative Predictive Value; *PPV*, Positive Predictive Value.

[1]The sample size showed in the table includes only the number of subjects used for the machine learning model development, with subjects used for other purposes, such as statistical analysis, not being included in this number.

a)

Authors	Sensitivity	2.5%	97.5%	Specificity	2.5%	97.5%
Delfin, 2019	0.750	0.498	0.886	0.845	0.674	0.935
Kirchebner, 2020a	0.830	0.777	0.873	0.739	0.651	0.811
Kirchebner, 2020b	0.826	0.780	0.865	0.801	0.700	0.875
Linaker, 1995	0.985	0.870	0.998	0.811	0.696	0.890
Pflueger, 2015	0.841	0.768	0.894	0.860	0.790	0.909
Sonnweber 2021	0.727	0.674	0.775	0.625	0.513	0.725
Thomas, 2005	0.545	0.415	0.673	0.823	0.795	0.850
Wang, 2020	0.630	0.534	0.716	0.321	0.256	0.394
Watts, 2021	0.873	0.785	0.961	0.605	0.459	0.751
AVERAGE	0.733	0.640	0.796	0.729	0.639	0.796
Test for equality of sensitivities: **X-squared** = 281.09, ***p*-value** = <0.000001						
Test for equality of specificities: **X-squared** = 382.63, ***p*-value** = <0.000001						
Correlation of sensitivities and false positive rates: **Rho** = 0.150 (-0.571-0.740)						
Total DOR: 9.57 (95% CI: 4.03-22.72), τ2=9.57 (95% CI: 0.00-6.93)						
Log DOR: 2.466 (95% CI: 1.534-3.397)						
posLR: 3.083 (95% CI: 1.954-4.866), τ2= 0.437 (0.000-0.947)						
negLR: 0.342 (95% CI: 0.201-0.583), τ2= 0.566 (0.000-0.3476)						
AUC: 0.816 (95% CI: 0.745-0.875); **pAUC**: 0.733						

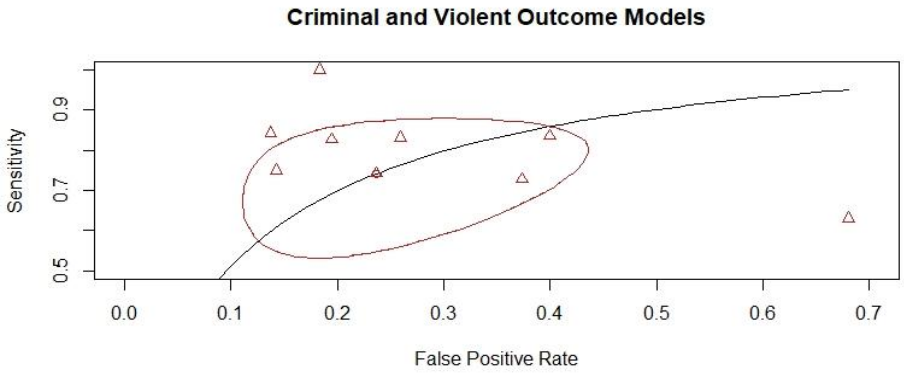
b)

Authors	mean	95% CI	%W(random)
Delfin, 2019	80.50	68.92-94.02	10.1
Kirchhebner, 2020	79.40	76.04-82.90	11.5
Kirchhebner, 2022	75.84	71.65-80.27	11.4
Linaker, 1995	90.65	83.78-98.07	11.2
Pflueger, 2015	85.00	80.62-89.60	11.4
Sonnweber, 2021	67.82	61.31-75.01	10.9
Thomas, 2005	57.50	51.20-64.57	10.7
Wang, 2020	47.00	46.53-47.46	11.6
Watts, 2021	71.58	67.04-76.42	11.3
AVERAGE	71.45	60.88-83.85	100%
Number of Observations: 2798, $\tau 2$= 0.042 (95% CI: 0.018-0.153)			

Table 2: Performance Metrics: Accuracies, AUC, diagnostic odds ratio, and likelihood ratios

A) Using the retisma function in mada, a linear mixed model with random effects was selected to produce summary estimates of sensitivity and specificity, as well as calculate AUC and partial AUC summary receiver operating characteristic (ROC) curves. Spearman's rho was used to assess correlation between sensitivities and false positive rates of included studies. The total diagnostic odds ratio (DOR), and positive and negative likelihood ratios (posLR, negLR) were calculated in a random effects model with a DerSimonian-Laird estimator using the maduani function in mada. The 95% confidence interval (CI) for AUC was calculated using bootstrapping with 2000 iterations with the dmetatools package in R. The average AUC across models was 0.816 (95% CI: 0.745-0.875), with a partial AUC of 0.733, and log DOR of 2.466 (95% CI: 1.534-3.397).

B) Using the metamean function in meta, the pooled accuracy of criminal and violent models was performed across 2428 patients (two studies used the same sample n=370) within a random-effects model using a restricted maximum likelihood estimator to calculate the variance $\tau 2$. Knapp-Hartung adjustments were used to calculate the confidence interval around the pooled effect. The average accuracy across models was 71.45% (95% CI: 60.88-83.86), with a heterogeneity variance $\tau 2$ of 0.0424.



Supplementary Figure S1: False Positive Rate Against Sensitivity Across Studies

Quality Scores of All Studies

CRIMINAL OUTCOMES										
Authors	Representative	Confounding	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Cohen, 1988	Yes	Yes	1)	No	Yes	No	Yes	No	Yes	6/9
Delfin, 2019	No	Yes	1)	Yes	Yes	Yes	Yes	Yes	No	7/9
Falconer, 2014	Yes	No	1)	Yes	Yes	No	No	No	Yes	5/9
Grann, 2007	Yes	No	1)	Yes	No	No	No	No	Yes	4/9
Kirchebner, 2020	No	Yes	2)	Yes	No*	No	Yes	Yes	Yes	6/9
McDermott, 2006	No	Yes	2)	No	Yes	No	No	Yes	No	3/9
Pflueger, 2015	Yes	No	2)	Yes	Yes	No	No	Yes	No	4/9
Sonnweber, 2021	No	Yes	2)	Yes	Yes	Yes	Yes	Yes	Yes	7/9
Watts, 2021	Yes	Yes	2)	Yes	Yes	Yes	Yes	Yes	Yes	8/9

VIOLENT OUTCOMES										
Authors	Representative	Confounding	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Kirchebner, 2022	No	Yes	2)	Yes	No	Yes	Yes	Yes	No	5/9
Le, 2018	Yes	No	2)	Yes	No	Yes	Yes	No	Yes	6/9
Linaker, 1996	No	No	2)	No	Yes	No	No	Yes	No	4/9
Menger, 2018	Yes	Yes	2)	Yes	Yes	No	Yes	No	No	6/9
Menger, 2019	Yes	Yes	2)	Yes	Yes	No	Yes	No	Yes	7/9
Monahan, 2000	Yes	Yes	2)	No	Yes	Yes	Yes	No	No	6/9
Steadman, 2000	Yes	Yes	2)	No	No	Yes	Yes	No	No	5/9
Suchting, 2018a	No	Yes	3)	Yes	Yes	Yes	Yes	Yes	No	6/9
Suchting, 2018b	Yes	Yes	2)	Yes	No	Yes	Yes	No	Yes	7/9
Thomas, 2005	Yes	Yes	2)	Yes	Yes	No	No	Yes	No	6/9

Tzeng, 2004	No	Yes	2)	Yes	No	No	Yes	No	No	3/9
Wang, 2020	No	Yes	2)	Yes	Yes	No	Yes	Yes	No	5/9

Supplementary Table S1: Quality of all studies

We created a machine learning quality assessment table based on experts' opinion to evaluate the reproducibility and reliability of the included studies. Our assessment provides a quick way to evaluate published papers and can also serve as a checklist for future studies. Briefly, the instrument comprises nine methodological considerations, including representativeness of the sample, confounding variables, outcome assessment, algorithm selection, feature selection, class imbalance (where applicable), missing data, performance/accuracy, and testing/validation. Further details can be found in the Supplementary Material.

* Kirchebner 2020: Feature selection was performed by ranking all variables, in order of importance, according to how often they were identified as top variables across backward selection, logistic regression, trees, SVMs and naive bayes. However, the exact way this was operationalized is unclear.

First author, year	Data utilized	Outcome	Sample size and diagnosis[1]	Validation	Machine learning model	Accuracy	Other measures
VIOLENT BEHAVIOR							
Barzman, 2018	Demographic variables, assessments of aggression, and static risk factors	Risk of school violence	103 middle and high school students recruited through outpatient clinics, inpatient units, and emergency department	Nested 10-fold cross-validation	LR with L2 normalization	N/A	91.02% (assessments only) 91.45% (assessments, clinical and sociodemographic data)
Gardner, 1996	Clinical record data	Violence was determined using incident reports from psychiatric, arrest, or criminal records and clinical interviews.	784 subjects with a psychiatric diagnosis (schizophrenia, affective disorders, substance use disorders, personality disorders, and others)	Not cross-validated	CART	N/A	Sensitivity / Specificity One-stage RT: 7.7% / 99.2% One-stage NBR: 9.3% / 99.1% Two-stage RT: 6.9% / 99.3% Two-stage NBR: 6.9% / 99.5%
Rosellini, 2018	Pre-Post Deployment Study (PPDS) of the Army STARRS dataset	Risk of interpersonal violence	7081 soldiers deployed to Afghanistan.	10-fold cross-validation; no external validation.	Ensemble learning: LR (EN with varying mixing parameter penalties, two SR, APS, two DT methods, BART, SVM, GBM, and NN)	N/A	Predictive models developed for each outcome, including depression (AUC 0.88), generalized anxiety disorder (AUC 0.85), suicidality (AUC 0.86) and head injury (AUC 0.74). Super learner AUC was 0.79 for anger attacks, 0.80 for being bullied or hazed, and 0.75 for getting into a fight. The sensitivity, specificity, and balanced accuracy of the models were not reported.
Thomas, 2005	Data from a large randomized controlled trial in 4 inner-city	Predictors of violence among patients with	780 patients with psychosis, 158 of	10-fold cross-validation; no	Full logistic regression (14	N/A	**Full logistic regression**

	mental health services in the United Kingdom (clinical/demographic variables)	psychosis	which were violent during the 2-year follow-up period.	external validation used.	variables) Forward stepwise logistic regression (6 variables) Full CART (123 nodes) Pruned CART (22 nodes) Pruned CART (22 nodes; violent cases given, 5 x weight)		Sensitivity - 19% Specificity - 96% PPV - 49% NPV - 79% Percent correctly classified - 77% **Forward Stepwise Logistic Regression** Sensitivity - 12% Specificity - 45% PPV - 41% NPV - 78% Percent correctly classified - 76% **Full CART** Sensitivity - 21% Specificity - 86% PPV - 31% NPV - 79% Percent correctly classified - 71% **Pruned CART** Sensitivity - 14% Specificity - 93% PPV - 38% NPV - 78% Percent correctly classified - 71% **Pruned CART** (22 nodes; violent cases

							given, 5 x weight) Sensitivity - 19% Specificity - 87% PPV - 30% NPV - 75% Percent correctly classified - 71%
CRIMINAL OUTCOMES							
Ang, 2013	Clinical questionnaires	Being charged or not charged for initial juvenile offending	2,899 adolescents from four school geographic areas	Holdout validation	LR DT ANN SVM	Testing / validation LR: 94.50 / 95.20 DT: 96.64 / 97.46 ANN: 97.22 / 98.26 SVM: 94.16 / 94.95	AUC LR: 0.950 DT: 0.968 ANN: 0.973 SVM: 0.946
Brodzinski, 1994	Clinical and demographic data	Differentiating criminal recidivists from non-recidivists	778 juvenile probation cases	Training (90%) and testing (10%) samples	Discriminant analysis ANN	63% (discriminant) 99% (ANN)	N/A
Caulkins, 1996	Clinical and administrative data	Criminal recidivism	3508 offenders during a two-year period following release from federal prison	Holdout validation with training (57.9%) and testing (41.9%) samples	LR MNN	Eighteen variable model: LR: 0.689 MNN: 0.699 Eleven variable model: LR: 0.683	N/A

						MNN: 0.689 Eight variable model: LR: 0.673 MNN: 0.684	
Cope, 2014	sMRI coupled with clinical assessments and sociodemographic data	Distinguishing homicide offenders from non-offenders	155 youth from a maximum-security facility	Two nested LOOCV	SVM with feature selection	81.29% (feature selection) 78.06% (no feature selection)	With feature selection: Specificity: 75.00% Sensitivity: 82.22% No feature selection: Specificity: 70.00% Sensitivity: 79.26%
Liu, 2011	HCR-20 questionnaire	Reconviction by violent offenses	882 male prisoners in England and Wales prospectively followed by a mean follow-up time of 3.31 years (1.34-4.24)	Holdout validation, with training (50%, testing (25%) and validation (25%) sets	LR CART ANN	N/A	Train LR: 0.72-0.75 CART: 0.67-0.71 MLPNN: 0.71-0.78 Test LR: 0.63-0.68 CART: 0.60-0.66 MLPNN: 0.65-0.70

							Validation LR: 0.64-0.66 CART: 0.58-0.66 MLPNN: 0.64-0.70
Palocsay, 2000	Nine clinical/demographic variables	Criminal recidivism among individuals released from prison	10357 prisoners in two cohorts	Holdout validation with training (n=2620), testing (n=7382) and validation sets (n=355)	Linear regression ANN	1978 ANN: 69.23% 1978 Logistic regression: 66.73% 1980 ANN: 66.98% 1980 Logistic regression: 65.71% 1978/1980 ANN: 65.96% 1978/1980 Logistic regression: 64.29	Recidivist correct (%) 1978 ANN: 41.26 1978 Logistic regression: 30.41 1980 ANN: 40.93 1980 Logistic regression: 30.53 1978/1980 ANN: 39.01 1978/1980 Logistic regression: 36.35 Non-recidivist correct (%) 1978 ANN: 85.89 1978 Logistic regression: 88.43 1980 ANN: 82.84 1978/1980 ANN: 82.15 1978/1980 Logistic regression: 81.07
Rosellini, 2016	Clinical and administrative data from the Army STARRS dataset	first accusation of a major physical violent crime	975 057 soldiers in the US Army in 2004–2009	Training (975, 057) and independent testing sample (43,248); of 10-fold cross-validated forward stepwise regression used for	Stepwise regression, random forests, penalized regressionS	0.80-0.82 AUC in the training dataset and 0.77 AUC in the validation dataset	Sensitivity, specificity, PPV and NPV were not reported. In the training dataset, an AUC of 0.81 was observed among men and 0.80-0.82 among women

				feature selection			
Rosellini, 2017	Clinical and administrative data Army STARRS dataset	Any crime with sufficient evidence to warrant an investigation	25,966 men and 2728 women who committed a first founded minor violent crime	10-fold cross-validation; external testing sample used	Stepwise and Penalized regression RF	AUC was 0.79 (for men and women) in the 2004-2009 training sample and 0.74-0.82 (men-women) in the 2011-2013 test sample.	N/A
Silver, 2000	Official clinical and administrative information of offenders convicted of an indictable offense	Risk of reimprisonment and rearrest following 1 year or 5 years after release	11749 offenders convicted of an indictable offense between October 1976-November 1977	Holdout validation with training (n=5875) and testing (n=5874) sets	LR CT Iterative LR ICT	Prison 1 year: 69.3%-83.7% Prison 5 years: 66.5%-82.5% Arrest 1 year: 45.6%-68.0% Arrest 5 years: 54.0%-82.2%	N/A
Silver, 2002	Official clinical and administrative information of offenders convicted of an indictable offense	Recidivism (imprisonment within 1 and 5 years, and arrest within 1 and 5 years	11749 offenders convicted of an indictable offense between October 1976-November 1977	Divided data into 10 subsamples, where 1 was used to construct the risk assessment model and 9 were used for cross-validation	LR ICT Feature selection using forward stepwise logistic regression	N/A	Model 1-10 Prison 1 year: 0.77-0.85 AUC Prison 5 years: 0.73-0.78 AUC Arrest 1 year: 0.73-0.76 AUC Arrest 5 years: 0.73-0.77 AUC Multiple Models - full

							Prison 1 year: 0.89 AUC Prison 5 years: 0.81 AUC Arrest 1 year: 0.79 AUC Arrest 5 years: 0.79 AUC Multiple Models - reduced Prison 1 year: 0.90 AUC Prison 5 years: 0.81 AUC Arrest 1 year: 0.78 AUC Arrest 5 years: 0.80 AUC
Stalans, 2004	Clinical and demographic variables obtained through clinical charts and legal records.	Violent recidivism while on probation	1344 violent offenders on probation	LOOCV; no external validation used	CTA - comparing against a logistic model with and without interaction	CTA 78.6% accuracy Logistic without interaction 81.8% accuracy Logistic with interaction 81.8% accuracy	CTA: sensitivity; 88.4% specificity Logistic without interaction: 9.8% sensitivity; 98.7% specificity Logistic with interaction :8.84% sensitivity; 98.9% specificity
Vilares, 2017	fMRI collected during a decision-making task	Mental states (knowledge and recklessness) when committing a hypothetical crime	40 healthy controls	Double-cross validation; no external validation used	Elastic-Net Regression	AUC of 0.792 average correct classification rate (CCR) of 71%	N/A

Haarsma, 2020	Tablet-based neuropsychological tests	Criminal recidivism	730 probationers	Training (80%) and testing (20%) samples	GLM LDA k-NN SVM GBM RF EN	N/A	Testing / validation (RFE model) GLM: 0.68 AUC LDA: 0.69 AUC k-NN: 0.60 AUC SVM (polynomial): 0.67 AUC GBM: 0.67 AUC RF: 0.66 AUC EN: 0.70 AUC
Delfin 2019	resting-state regional cerebral blood flow (rCBF) and clinical risk factors	Criminal recidivism	44 forensic psychiatry patients	Out-of-bag (OOB) error	RF	Accuracy: 82% Sensitivity: 75% Specificity: 86%	AUC: 0.81 PPV = 0.73 NPV = 0.86 Note: the dataset was not split into training and testing sets, and OOB error was used as a resampling procedure
OTHER OUTCOMES							
Monaro, 2018	Behavioral Measures (mouse-movements during a computerized task)	Malingering of clinical depression	100 individuals both with and without clinical depression	Holdout validation with training (n=60 and test(n=27) sets	NB SMO LMT RF Feature selection: Correction based feature selection	<u>Accuracy in 10-fold-cross validation (n=60)</u> Naive Bayes - 80 SMO - 82.5 LMT - 80 Random Forest - 87.5 <u>Accuracy in test set</u>	Note: authors did not report sensitivity, specificity, PPV or NPV

					Validation: 10-fold cross validation	(n=28) Naive Bayes - 94.4 SMO - 88.9 LMT - 88.9 Random Forest - 94.4	
Ponseti, 2012	fMRI blood oxygen level-dependent signals to child and adult sexual stimuli for each participant	Identification of pedophilia	24 participants with pedophilia 32 healthy controls	LOOCV	LDA k-NN	k-NN: 75-91% LDA: 89-95%	Sensitivity / Specificity k-NN: 63-88% / 84-94% LDA: 88-92% / 88-100%
Ponseti, 2015	Haemodynamic fMRI response to face images of women, girls, men, and boys	Classification of Pedophilia	24 males diagnosed with pedophilia according to the DSM-IV-R (11 heterosexual pedophiles, 13 homosexual pedophiles).	LOOCV; no external validation	Fisher's linear discriminant analysis	Mean classification accuracy of 93%	91% specificity 95% sensitivity
Rosenfeld, 2005	Official clinical and sociodemographic records from criminal defendants	Stalking behavior	204 individuals evaluated for crimes related to stalking or obsessional harassment	Jack-knife classification approach of training sample	CART models comprising: Tree regression, Logistic regression,	N/A	Tree regression - AUC .649 Logistic regression - AUC .706

Mazza, 2018	Computerized neuropsychological test	Malingering	175 individuals	10-fold CV Hold-out validation with training (70.6%) and testing (29.4%) sets	LR SVM NB RF LMT	Time-pressure models: 95% accuracy across all models Non time-pressure models: Accuracy ranged from 75-95%	AUC not reported Note: it is important to mention that a small testing set was used (n=20), which may yield inflated accuracy.
Pace, 2019	Test taking effort assessment (b Test)	Malingering	63 individuals	LOOCV	NB LR SL SVM RF	NB: 90.47% LR: 90.47% SL: 92.9% SVM: 88.09% RF: 90.47%	NB: 0.89 AUC LR: 0.85 AUC SL: 0.91 AUC SVM: 0.88 AUC RF: 0.89 AUC Note: sensitivity and specificity not reported. The model also did not separate the data into training and testing sets, as such, model accuracy may be inflated.

Supplementary Table S2 – Machine learning studies predicting criminal and violent outcomes in non-psychiatric individuals.

Authors	Classification Task	Method to address class imbalance	True and False Positive/Negative	Performance Metrics	95% Confidence Intervals of Accuracy
Delfin 2019	Recidivists (N=16) Non-recidivists (N=28)	Downsampling of majority class	TP = 12 FP = 4 FN = 4 TN = 24	Balanced Accuracy = 80.5% Sensitivity = 75% Specificity = 86% False Positive = 14% False Negative = 25% Standard Error = 6.3775	Accuracy: 80.5% (95% CI: 68.92-94.02%)
Kirchhebner, 2020	Recidivists (N=209) Non-recidivists (N=135)	None	TP = 193 FP = 29 FN = 39 TN = 83	Balanced Accuracy = 79.4% Sensitivity = 83% Specificity = 74% False Positive = 26% False Negative = 17% Standard Error = 2.2168	Accuracy = 79.4% (95% CI: 76.04-82.91%)
Kirchhebner, 2022	Violent offenders (N=294) Non-violent offenders (N=75)	SMOTE	TP = 254 FP = 15 FN = 53 TN = 62	Balanced Accuracy = 75.84% Sensitivity = 80.49% Specificity = 71.19% False Positive = 28.81% False Negative = 19.51% Standard Error = 2.1989	Accuracy = 75.84% (95% CI: 71.65-76.43%)
Linaker, 1995	Violent patients (N=32) Non-violent patients	None	TP = 32 FP = 11	Balanced Accuracy = 90.65% Sensitivity = 100%	Accuracy = 90.65% (95% CI: 83.79-98.07%)

	(N=60)		FN = 0 TN = 49	Specificity = 81.3% False Positive = 18.7% False Negative = 0% Standard Error = 3.6403	
Pflueger, 2015	Recidivists (N=128) Non-recidivists (N=131)	None	TP = 108 FP = 18 FN = 20 TN = 113	Balanced Accuracy = 85% Sensitivity = 84% Specificity = 86% False Positive = 14% False Negative = 16% Standard Error = 2.2908	Accuracy = 85.00% (95% CI: 80.64-89.61%)
Sonnweber, 2021	Violent offenders (N=294) Non-violent offenders (N=75)	Oversampling minority class	TP = 214 FP = 28 FN = 80 TN = 47	Balanced Accuracy = 67.82% Sensitivity = 72.73% Specificity = 62.92% False Positive = 37.08% False Negative = 27.27% Standard Error = 3.4872	Accuracy = 67.82% (95% CI: 61.32-75.01%)
Thomas, 2005	Violent at follow-up (N=158) Non-violent at follow-up (N=622)	None	TP = 30 FP = 25 FN = 128 TN = 597	Balanced Accuracy = 57.5% Sensitivity = 19% Specificity = 96% False Positive = 4% False Negative = 81%	Accuracy = 57.50% (95% CI: 51.20-64.57%)

				Standard Error = 1.8035	
Wang, 2020	Violent at follow-up (N=103) Non-violent at follow-up (N=172)	None	TP = 65 FP = 117 FN = 38 TN = 55	Balanced Accuracy = 47% Sensitivity = 63% Specificity = 32% False Positive = 68% False Negative = 37% Standard Error = 0.4	Accuracy = 47% (95% CI: 46.54-47.47%)
Watts, 2021	Violent & Non-violent offenses (N=1116) Sexual offenses (N=124) *Following downsampling*: Violent & Non-violent offenses (N=248) Sexual offences (N=74)	Downsampling of the majority class	*Metrics following downsampling*: TP = 61 FP = 99 FN = 12 TN = 149	Balanced Accuracy = 71.58% Sensitivity = 83.15% Specificity = 60.00% False Positive = 40% False Negative = 16.85% Standard Error = 2.3928	Accuracy = 71.58% (95% CI: 67.04-76.43%)

Supplementary Table S3: Confusion Matrices of Classification Models

False positive rate is calculated as 1-specificity, while false negative is calculated as 1-sensitivity. Standard error was calculated by subtracting the upper bound of the 95% CI from the lower bound and dividing by 3.92 (upper bound - lower bound)/3.92. This standard error calculation was used for all studies, apart from Wang et al. 2020, which reported standard error as 0.4. 95% confidence intervals are reported as calculated using an inverse variance method within a random effects model. Additionally, confusion matrices were provided according to the method used to address class imbalance, where applicable. It is important to note that none of the included studies reported the true positives/true negatives and false positives/false negative rates, and the numbers indicated in the table reflect calculations based on the prevalence, sensitivity, specificity, and total sample size.

SUPPLEMENTARY MATERIAL

Scopus
((artificial AND intelligence) OR (supervised AND machine AND learning) OR (k-nearest AND neighbors) OR (decision AND trees) OR (naive AND bayes) OR (random AND forest) OR (gradient AND boosting) OR (elastic AND net) OR (support AND vector AND machine) OR (relevance AND vector AND machine) OR (Latent Class Analysis) OR (Neural Networks)) AND ((commitment AND of AND mentally AND ill) OR (insanity AND defense)) OR ((criminals) OR (schizophrenia) OR (schizophrenia AND spectrum AND other AND psychotic AND disorders) OR (psychotic AND disorders) OR (forensic AND psychiatry))

Results: 6531
Search Date: 2022-04-18

PubMed
((((((((("Artificial Intelligence/classification"[Mesh] OR "Artificial Intelligence/methods"[Mesh])) AND ("Supervised Machine Learning/classification"[Mesh]) OR "Commitment of Mentally Ill/statistics and numerical data"[Mesh])) OR ("Insanity Defense/classification"[Mesh] OR "Insanity Defense/statistics and numerical data"[Mesh])) AND ("Criminals/classification"[Mesh] OR "Criminals/statistics and numerical data"[Mesh])) AND ("Schizophrenia/classification"[Mesh] OR "Schizophrenia/diagnostic imaging"[Mesh] OR "Schizophrenia/statistics and numerical data"[Mesh])) AND ("Schizophrenia Spectrum and Other Psychotic Disorders/classification"[Mesh] OR "Schizophrenia Spectrum and Other Psychotic Disorders/diagnosis"[Mesh] OR "Schizophrenia Spectrum and Other Psychotic Disorders/statistics and numerical data"[Mesh])) AND ("Psychotic Disorders/diagnosis"[Mesh] OR "Psychotic Disorders/statistics and numerical data"[Mesh])) OR ("Forensic Psychiatry/classification"[Mesh] OR "Forensic Psychiatry/statistics and numerical data"[Mesh]))

Results: 1613
Search Date: 2022-04-18

Web of Science
(TS=(Artificial Intelligence) OR TS = (Supervised Machine Learning) OR TS = (Deep Learning) OR AB = (Support Vector Machin*) OR AB = (Relevance Vector Machin*) OR AB = (Random Forest) OR AB = (Decision Tree*) OR AB = (Gradient Boost*) OR AB = (Extreme Boost*) OR AB = (Elastic Net) OR AB = (Logistic Regression) OR AB = (Naive Bayes) OR AB= (Neural Network*) OR TS = (Expert System*) OR TS = (Latent Class Analys*)) AND (TS = (Forensic Psych*) OR TS = (Commitment of Mentally Ill) OR TS = (Insanity Defence) OR TS = (Crimin*) OR TS = (Offend*))

Results: 4792
Search Date: 2022-04-18

Total records(before duplicate removal): 12936

Total records(duplicates removed): 12420

2. Quality assessment instrument development

We formed a group of multidisciplinary researchers from the fields of Neuroscience, Psychiatry, and Computer Science to develop a time efficient and practical assessment strategy to evaluate the quality of machine learning based healthcare research. For that purpose, we attempted to capture the reliability of the results presented in a given study and identify practical ways that methodology may be improved.

This comprised nine methodological features, including sample representativeness, confounding variables, and outcome assessments, which were judged to be the most clinically pertinent components in machine learning-based healthcare research. Relevant considerations of each methodological feature are discussed in further detail in the next sections. The six remaining dimensions assess the quality and specific components of the machine learning approach that were used in a given study. In summary, this entails the algorithm or framework used, evidence that hyper-parameter optimization and feature selection procedures were used, whether authors provided details on how missing data and class imbalance problems were handled, the accuracy of a given model, and finally whether the model performance was tested in unseen data. These dimensions were qualitatively evaluated according to the information in section 3.

3. Quality assessment instrument domains

Methodological Feature	Considerations
1. Representativeness of the sample	Was the study representative of the heterogeneity observed in the target population? If not, was this related to the sampling method, insufficient sample size or inclusion/exclusion criteria?
2. Confounding variables	Did the study control for the most relevant confounding variables? If so, were covariates assessed using subjective or objective measures?
3. Outcome assessment	How were outcome measures assessed: A. Independent blind assessment (✓) B. Secure record (e.g., surgical records) (✓) C. Interview not blinded, self-report or medical record D. No description
4. Algorithm selection	Was the machine learning algorithm used to analyse the data clearly described and appropriate?
5. Feature selection	Did the study describe both feature selection and hyperparameter tuning? Which metrics were used?
6. Class imbalance	Did the authors address the class imbalance problem? Which method was used?
7. Missing data	Did the study describe how the authors handled missing data, including whether they were inputted or removed?
8. Performance/accuracy	Were the following performance metrics included for classification studies? A. Accuracy B. Sensitivity

	C. Specificity D. AUC E. PPV/NPV F. 95% Confidence intervals of performance metrics Or, alternatively, were one of the following performance metrics included for regression studies? A. Mean-squared error B. Mean-absolute error C. Root-mean-squared error
9. Testing/validation	Was the test dataset "unseen" regarding model training? Was the model tested on a hold-out or an external dataset?

3.1. Representativeness of the sample

Machine learning models can deal with large amounts of data and the problem of heterogeneity. Therefore, there is less of a need to be restrictive with inclusion and exclusion criteria. Here, we evaluated whether the sample selected by the authors reflected the real population being studied. When the sample did not reflect the population being studied, we evaluated if it was because (1) the sampling methods were not appropriate, (2) the sample was not large enough to represent the population or (3) the inclusion and exclusion criteria restricted the individuals in the study.

3.2. Confounding Variables

To adequately control for confounding variables in machine learning, we need to ensure that they will have a similar effect across the entirety of the sample. To achieve this, randomization is used throughout the analysis. More specifically, training and testing datasets are randomised using resampling techniques, and the analysis is often repeated with different parameters and learning decisions (parameter tuning). Using the criteria, we evaluated whether the authors controlled for confounding variables.

3.3. Outcome assessment

How an outcome is defined has several important implications in a predictive model. Depending on the question or problem, a classification task may be appropriate, which uses a categorical outcome, or a regression task may be more relevant, which has a continuous numeric outcome. A clinical instrument or questionnaire, for example, can be used as a numeric score or it can be transformed into a categorical outcome by using a cut-off. We evaluated how authors assessed these outcomes, considering (1) independent blind assessments and secure records as high quality, (2) unblinded interview, self-report or medical record as lower quality and (3) when no description was available.

3.4. Algorithm selection

There are several algorithms to choose from, with each relying on slightly different assumptions of the underlying data. Broadly speaking, there are linear (logistic regression, linear support vector machine), non-linear (Naive Bayes, K-Nearest Neighbours, Learning Vector Quantization), tree-based (decision trees, random forest, xgboost) and neural network (convolutional neural network, multilayer perceptrons) models, although others exist. Certain algorithms may be better suited to particular problems. For example, tree-based models such as random forest may be better suited to datasets with multicollinearity

among features than linear-based models such as logistic regression. However, regularisation parameters can be used in linear-based models (such as L2 regularisation) to account for issues such as this. Nevertheless, it is often difficult to determine beforehand which algorithms will lead to the highest model performance. Therefore, it is often a good strategy to compare the model performance of several algorithms. In this item, we evaluated whether the authors used an algorithm that is commonly used for the specific type of dataset, if several algorithms were compared, and if hyperparameter tuning was used.

The appropriateness of a machine learning algorithm was determined based on whether the specific data used in model development was congruent or incongruent with the strengths and limitations of the specific algorithm. For example, if a Gaussian process model was used, which is a non-sparse algorithm that loses efficiency in high dimensional spaces, in conjunction with a high-dimensional dataset, this algorithm would be deemed inappropriate for the input data. Conversely, Naive Bayes, which works well with high dimensional data would be considered an appropriate algorithm in such cases. Another example of an inappropriate model would be the use of convolutional neural networks for structural and tabular style datasets, as such algorithms are better suited to unstructured datasets. In cases where authors included both appropriate and inappropriate algorithms during model development, this consideration is scored with a "B", alongside an asterisk to indicate which algorithms were inappropriate and why. Studies which only utilised one algorithm during model development that was deemed inappropriate received a score of "C". Furthermore, studies are scored with a "B" if they did not compare multiple algorithms during model development and were scored as an "A" if they compared multiple algorithms that were deemed appropriate based on the candidate feature set.

3.5. Feature selection

A common problem in machine learning studies is the so-called small-n-large-p problem, also known as the curse of dimensionality, which occurs when there are more variables than examples in a dataset. Machine learning models created using these datasets are more prone to overfitting, which often results in overinflated performance in a training dataset, but much poorer performance in an external testing dataset. In addition, some algorithms cannot deal with more dimensions than examples. Highly correlated variables can also introduce more importance to a specific characteristic, decreasing the importance of the remaining variables. To circumvent these issues, a proper feature selection procedure, when applicable, should be done prior to training or as part of the training procedure, such as it happens in embedded methods. The feature selection can be knowledge-driven or data-driven. In this item, we examined if the study used a proper feature selection (if applicable).

3.6. Class imbalance

Class imbalance occurs when the distribution of the outcome classes is highly unbalanced, i.e., when one outcome occurs much more frequently than the other one. This may result in a model with high accuracy but with very little clinical utility. For example, let us suppose that we have 99 occurrences of non-violence in our dataset and only 1 occurrence of a violent incident. Even if our model has 99% accuracy, it is useless if the model cannot detect the one violent incident with high accuracy. In this item, we evaluated whether there was a class imbalance in the sample and if this problem was correctly addressed. This can be done using a series of methods, including (1) changing the metric of performance (accuracy, for example, is a poor form of evaluating imbalanced data sets; (2) resampling the data set by artificially increasing it (oversampling) or by removing examples from the majority class to create a more balanced data set (undersampling); (3) by generating more data with algorithms such as the Synthetic Minority Over-Sampling Technique (SMOTE); (4) by choosing algorithms that deal better with unbalanced classes, such as CART or random forests; (5) by using penalised models; or (6) by using anomaly and change detection.

3.7. Missing data

It is critical to handle missing data since several algorithms cannot process incomplete data sets. Furthermore, it is also necessary to use an adequate imputation method to avoid introducing bias, which would otherwise lead to false conclusions if not addressed. It is important to report the amount of missing data in each variable, if these cases were excluded, or if the authors used an algorithm to input data and which algorithm/technique was used. Ideally, authors should provide a visual distribution of the patterns of missing data, such as aggregation plots, spinogram/spineplots, mosaic plots, etc. All these factors were evaluated in this section.

3.8. Performance/accuracy

Here, we evaluate whether the authors reported all relevant results and if they used the appropriate metrics. Studies informing only partial metrics may mask bias and flaws of the method, preventing the reader from fully understanding the relevance of the model.

3.9. Testing/Validation

We can divide the machine learning process into three main components: training, validation, and testing. A training set allows the algorithm to learn and develop a predictive model. The validation set contains unseen data and is used to control for overfitting. Frequently, the same dataset is divided into training and validation sets. After a model is trained and validated, and shows consistent performance in both these steps, the model can be applied in an external and independent testing set. This allows us to see if the model can be generalised outside of the original sample. Some validation methods include holdout validation, k-fold, and leave one out cross validation.

A model that shows good performance in the training set but performs significantly poorer in the validation step is most likely due to overfitting - which occurs when the model relies more on the specific nuances and noise of the training dataset, resulting in poor accuracy in unseen data. In this item, we evaluated whether the authors properly tested and validated their models by taking steps to improve its generalizability. It is important to highlight that the use of cross-validation to evaluate performance should be discouraged when the data is large enough for a training-test split. Furthermore, the size of the test set should be sufficiently large for accuracy and other metrics to be estimated with high reliability.

References

1. Faay, M. D. M. & Sommer, I. E. Risk and Prevention of Aggression in Patients with Psychotic Disorders. *American Journal of Psychiatry* **178**, 218–220 (2021).
2. Fazel, S., Wolf, A., Palm, C. & Lichtenstein, P. Violent crime, suicide, and premature mortality in patients with schizophrenia and related disorders: a 38-year total population study in Sweden. *The Lancet Psychiatry* **1**, 44–54 (2014).
3. Whiting, D., Gulati, G., Geddes, J. R. & Fazel, S. Association of Schizophrenia Spectrum Disorders and Violence Perpetration in Adults and Adolescents From 15 Countries. *JAMA Psychiatry* **79**, 120 (2022).
4. Douglas, K. S., Guy, L. S. & Hart, S. D. Psychosis as a risk factor for violence to others: A meta-analysis. *Psychological Bulletin* **135**, 679–706 (2009).
5. Penn, D. L., Kommana, S., Mansfield, M. & Link, B. G. Dispelling the Stigma of Schizophrenia: II. The Impact of Information on Dangerousness. *Schizophrenia Bulletin* **25**, 437–446 (1999).
6. Buchanan, A., Sint, K., Swanson, J. & Rosenheck, R. Correlates of Future Violence in People Being Treated for Schizophrenia. *American Journal of Psychiatry* **176**, 694–701 (2019).
7. Moulin, V. *et al.* Impulsivity in early psychosis: A complex link with violent behaviour and a target for intervention. *European Psychiatry* **49**, 30–36 (2018).
8. Storvestre, G. B. *et al.* Childhood Trauma in Persons With Schizophrenia and a History of Interpersonal Violence. *Frontiers in Psychiatry* **11**, (2020).
9. Keers, R., Ullrich, S., DeStavola, B. L. & Coid, J. W. Association of Violence With Emergence of Persecutory Delusions in Untreated Schizophrenia. *American Journal of Psychiatry* **171**, 332–339 (2014).
10. Swanson, J. W. *et al.* A National Study of Violent Behavior in Persons With Schizophrenia. *Archives of General Psychiatry* **63**, 490 (2006).
11. de Girolamo, G. *et al.* A multinational case−control study comparing forensic and non-forensic patients with schizophrenia spectrum disorders: the EU-VIORMED project. *Psychological Medicine* 1–11 (2021) doi:10.1017/S0033291721003433.
12. Whiting, D., Lichtenstein, P. & Fazel, S. Violence and mental disorders: a structured review of associations by individual diagnoses, risk factors, and risk assessment. *The Lancet Psychiatry* **8**, 150–161 (2021).
13. Fazel, S., Wolf, A., Palm, C. & Lichtenstein, P. Violent crime, suicide, and premature mortality in patients with schizophrenia and related disorders: a 38-year total population study in Sweden. *The Lancet Psychiatry* **1**, 44–54 (2014).
14. Sariaslan, A., Larsson, H. & Fazel, S. Genetic and environmental determinants of violence risk in psychotic disorders: a multivariate quantitative genetic study of 1.8 million Swedish twins and siblings. *Molecular Psychiatry* **21**, 1251–1256 (2016).

15. Fleischman, A., Werbeloff, N., Yoffe, R., Davidson, M. & Weiser, M. Schizophrenia and violent crime: a population-based study. *Psychological Medicine* **44**, 3051–3057 (2014).

16. Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. Criminal Justice and Behavior, 37(9), 965–988. https://doi.org/10.1177/0093854810374274

17. Kröner, C., Stadtland, C., Eidt, M. and Nedopil, N., 2007. The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. Criminal Behaviour and Mental Health, 17(2), pp.89-100.

18. Singh, J.P., Serper, M., Reinharth, J. and Fazel, S., 2011. Structured assessment of violence risk in schizophrenia and other psychiatric disorders: a systematic review of the validity, reliability, and item content of 10 available instruments. Schizophrenia bulletin, 37(5), pp.899-912.

19. Michel, S.F., Riaz, M., Webster, C., Hart, S.D., Levander, S., Müller-Isberner, R., Tiihonen, J., Repo-Tiihonen, E., Tuninger, E. and Hodgins, S., 2013. Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: Accuracy of prediction, general and forensic settings, and dynamic risk factors. International Journal of Forensic Mental Health, 12(1), pp.1-13.

20. Lobo, J.M., Jiménez-Valverde, A. and Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), pp.145-151.

21 Wiens, J. & Shenoy, E. S. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases* **66**, 149–153 (2018).

22. Cearns, M., Hahn, T. & Baune, B. T. Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry* **9**, 271 (2019).

23. *DIAGNOSTIC AND STATISTICAL MANUAL OF DSM-5 ™*.

24. Mullally, K., Mamak, M. & Chaimowitz, G. A. The next generation of risk assessment and management. *International Journal of Risk and Recovery* **1**, 21–26 (2018).

25. Cook, A. N. *et al.* Validating the Hamilton Anatomy of Risk Management–Forensic Version and the Aggressive Incidents Scale. *Assessment* **25**, 432–445 (2018).

26. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* (2008) doi:10.18637/jss.v028.i05.

27. Kuhn, M. caret Package. *Journal Of Statistical Software* (2008).

28. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* (2002).

29. Schonlau RAND, M. *Boosted regression (boosting): An introductory tutorial and a Stata plugin*. *The Stata Journal* vol. 5 (2005).

30. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320 (2005).

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

31. Zhang, Z. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine* **4**, 218–218 (2016).

32. Freund, Y. & Schapire, R. E. *A Short Introduction to Boosting*. *Journal of Japanese Society for Artificial Intelligence* vol. 14 www.research.att.com/ (1999).

33. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. (2016) doi:10.1145/2939672.2939785.

34. Breiman, L. Random forests. *Machine Learning* (2001) doi:10.1023/A:1010933404324.

35. Sutton, C. D. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics* vol. 24 303–329 Preprint at https://doi.org/10.1016/S0169-7161(04)24011-1 (2005).

36. Nasejje, J. B., Mwambi, H., Dheda, K. & Lesosky, M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology* **17**, 115 (2017).

37. Dash, M. & Liu, H. Feature selection for classification. *Intelligent Data Analysis* (1997) doi:10.3233/IDA-1997-1302.

38. Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. in *Data Classification: Algorithms and Applications* (2014). doi:10.1201/b17320.

39. Jovic, A., Brkic, K. & Bogunovic, N. A review of feature selection methods with applications. in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 1200–1205 (IEEE, 2015). doi:10.1109/MIPRO.2015.7160458.

42. Longadge, M. R., Snehlata, M., Dongre, S. & Latesh Malik, D. *Class Imbalance Problem in Data Mining: Review*. *International Journal of Computer Science and Network* vol. 2 www.ijcsn.org (2013).

43. Wong, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* **48**, 2839–2846 (2015).

44. Tharwat, A. Classification assessment methods. *Applied Computing and Informatics* **17**, 168–192 (2021).

45. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**, 197–227 (2016).

46. Adnan, N., Ahmad, M. H. & Adnan, R. *A Comparative Study On Some Methods For Handling Multicollinearity Problems*. *MATEMATIKA* vol. 22 (2006).

47. Beretta, L. & Santaniello, A. Nearest neighbour imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making* **16**, 74 (2016).

48. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research* **20**, 40–49 (2011).

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

49. Poulos, J. & Valle, R. *MISSING DATA IMPUTATION FOR SUPERVISED LEARNING †*. (2018).

50. Watts, D. *et al*. Predicting offences among individuals with psychiatric disorders - A machine learning approach. *Journal of Psychiatric Research* **138**, 146–154 (2021).

Chapter 4 - Stigmatized individuals: a case for precision ethics

Authors: Devon Watts[1,2]; Jeff D'Souza[3]; Marco Antonio Azevedo[4]; Flavio Kapczinski[1,2,5,6], Gary Chaimowitz[1,6]

1. St. Joseph's Healthcare Hamilton, Ontario, Canada
2. Neuroscience Graduate Program, McMaster University, Hamilton, Canada.
3. Institute on Ethics & Policy for Innovation, McMaster University, Hamilton, Canada.
4. Department of Philosophy, Universidade do Vale do Rio, Rio Grande do Sul, Brazil.
5. Instituto Nacional de Ciência e Tecnologia Translacional em Medicina (INCT-TM), Porto Alegre, Brazil.
6. Department of Psychiatry and Behavioral Neurosciences, McMaster University, Hamilton, Canada.

*Corresponding author:

Flavio Kapczinski, MSc, MD, PhD, FRCPC

Professor, Department of Psychiatry & Behavioural Neurosciences

Director, Centre for Clinical Neurosciences, McMaster University

Director, Neuroscience Graduate Program, McMaster University

100 West 5th Street, Hamilton, Ontario, L9C 0E3, Canada

Phone: 905-522-1155 Ext. 35420

Email: kapczinf@mcmaster.ca

Email for all authors:

Devon Watts: wattsd@mcmaster.ca

Jeffrey D'Souza: dsouzjj@mcmaster.ca

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

Marco Azevedo: mazevedogtalk@gmail.com

Flavio Kapczinski: kapczinf@mcmaster.ca

Gary Chaimowitz: chaimow@mcmaster.ca

This chapter in its entirety has been ***published*** in the journal **Trends Psychiatry Psychotherapy**. The final accepted manuscript version of this article is presented within this thesis.

Watts D, D'Souza J, Azevedo MA, Kapczinski F, Chaimowitz G. Stigmatized individuals: a case for precision ethics. Trends Psychiatry Psychother. 2021 Sep 2. doi: 10.47626/2237-6089-2021-0354. Epub ahead of print. PMID: 34551242.

Emerging technologies have enabled us to create increasingly accurate predictions about the propensity of psychiatric patients to commit criminal offenses.[1] Machine learning models raise a variety of opportunities and avenues to develop educational tools, preventive measures, and shape public policy.[2] However, despite the promise of predictive algorithms in forensic psychiatry, their use raises an important ethical challenge. Namely, how can we avoid further stigmatizing vulnerable individuals, and instead, ensure our algorithms respect their rights, enhance their safety, and promote their wellbeing? The noted philosopher Joel Feinberg envisioned a form of noncomparative justice, where each person is treated precisely as they deserve, without regard to the way anyone else is treated.[3]

To better elucidate this concept, take the example of "voluntary" or "involuntary" criminal acts, which depend on an individual's intention to commit a crime, otherwise known as *means rea* (guilty mind). When voluntary criminals are compared against voluntary criminals, such a system is thought to be fair and just in a legal sense. However, when involuntary criminals are compared with voluntary criminals in the same category, and are punished with similar severity, we can discern a state of injustice because of a difference in criminal culpability. As such, the voluntary nature of the criminal act, regardless of the severity of the crime, is a salient consideration.[4]

In many countries, individuals with severe mental illness who commit criminal acts are evaluated according to noncomparative justice.[5] Rather than simply punishing the offender in proportion to the severity and context of the crime, those with severe mental illness who lack *means rea* may be treated in a *restorative* framework, recognizing the need to aid, treatment, and seek to prevent future reoffending.[5] In forensic psychiatry, this implies the need for targeted and individualized treatment.

However, several pertinent questions arise when evaluating the utility and implementation of such algorithms. For instance, an important consideration that is often overlooked is model interpretability. So called "black box" methods may perform well in testing and validation datasets, however without a rudimentary understanding of the directionality, and interaction effects, of important features, we lack the transparency required to justify implementing these models in high stakes clinical settings.[6] Toward this end, new methods leveraging the internal structure of tree based algorithms can be used to directly measure local feature interaction effects, and provide insight into the magnitude, prevalence, and direction of a feature's effect.[7]

Similarly, even among classification models that demonstrate high accuracy, there will be instances where individuals are misclassified. In cases where the risks of misclassification are low, this may be largely unimportant. However, when dealing with the complex intersectionality between healthcare, personal freedom, and societal risk, this becomes a challenging consideration. For instance, how can we introduce ethical constraints in our models without significantly impacting their overall accuracy and utility? While this remains open to debate, it may be useful to consider such ethical goals from two distinct frameworks.

Robert Nozick, the renowned American philosopher, once discussed the concept of moral pushes and pulls.[8] *Moral pushes* involve ideals or values that propel us "from within". From this framework, ethics are a set of principles that help guide us to being more virtuous individuals. Ethical algorithms can favour these individual moral values if the goal is to make us "better people", allowing us to live a healthier life, or intrinsically, boosting moral dispositions so that we can better operate within society, leading to the benefit of others by proxy. *Moral pulls*, on the other hand, are constraints about the design of the algorithms. For instance, ensuring that our models are not predicated on immutable characteristics, and ensuring free, informed, and

ongoing consent.[8] The concept of moral pulls also highlights the importance of patient centred perspectives. We argue that a prerequisite for the successful implementation of predictive models into routine care is for data scientists to meaningfully engage with stakeholders (healthcare providers, patients, and their families) to ensure the scope of the problem, and important ethical considerations, are adequately elucidated.

Altogether, we advocate for a marked transformation in the field, where group level statistical approaches to risk assessment, therapeutic interventions, and rehabilitation are abandoned in favour of more precise, individualized models, developed according to a new, precision ethics approach.

Conflict of interest statements:

FK has received grants or research support from AstraZeneca, Eli Lilly, Janssen-Cilag, Servier, NARSAD, and the Stanley Medical Research Institute; has been a member of speakers' boards for AstraZeneca, Eli Lilly, Janssen and Servier; and has served as a consultant for Servier. The other authors declare no competing interests.

Author's contributions:

All authors participated in the writing, revisions, and the approval of the final manuscript.

Role of funding source:

Not applicable

Ethics committee approval:

Not applicable

References

1. Watts D, Moulden H, Mamak M, Upfold C, Chaimowitz G. Predicting offenses among individuals with psychiatric disorders - A machine learning approach. J Psychiatr Res. 2021;138(October 2020):146–54.
2. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. The Lancet Psychiatry. 2016.
3. Feinberg J. Philosophical Review. 2014;83(3):297–338.
4. Gerber RJ. Insanity and Mens Rea. In: Insanity Defense. Associated Faculty Press; 1984. p. 98–117.
5. Naude B. An international perspective of restorative justice practices and research outcomes. J Juridical Sci. 2006;31(1):101–20.
6. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;
7. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;
8. Nozick R. Philosophical explanations. Harvard University Press; 1981.

Chapter 5 - Intranasal esketamine and the dawn of precision psychiatry

Authors: Devon Watts[1]; Frederico D. Garcia[2], Acioly L.T. Lacerda[3], Jair de J. Mari[3], Lucas C. Quarantini, Flavio Kapczinski

St. Joseph's Healthcare Hamilton, Ontario, Canada

7. Neuroscience Graduate Program, McMaster University, Hamilton, Canada.
8. Deparamento de Saude Mental, Universidade Federal de Minas Gerias (UFMG), Belo Horizonte, MG, Brazil
9. Programa de Transtornos Afetivos (PRODAF), Departamento de Psiquitaria, Universidade Federal de Sao Paulo (UNIFESP), Sao Paulo, SP Brazil
10. Departamento de Neurociencias e Saude Mental, Faculdade de Medicina da Bahia, UFBA, Salvador, BA, Brazil
11. Professor, Department of Psychiatry & Behavioural Neurosciences; Director, Centre for Clinical Neurosciences, McMaster University; Director, Neuroscience Graduate Program, McMaster University

*Corresponding author:

Flavio Kapczinski, MSc, MD, PhD, FRCPC

Professor, Department of Psychiatry & Behavioural Neurosciences

Director, Centre for Clinical Neurosciences, McMaster University

Director, Neuroscience Graduate Program, McMaster University

100 West 5th Street, Hamilton, Ontario, L9C 0E3, Canada

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

Phone: 905-522-1155 Ext. 35420

Email: kapczinf@mcmaster.ca

This chapter in its entirety has been ***published*** in the Brazilian Journal of Psychiatry. The final accepted manuscript version of this article is presented within this thesis.

Watts D, Garcia FD, Lacerda ALT, Mari JJ, Quarantini LC, Kapczinski F. Intranasal esketamine and the dawn of precision psychiatry. Braz J Psychiatry. 2022 Mar-Abr;44(2):117-118. doi: 10.1590/1516-4446-2021-0031. PMID: 34320126; PMCID: PMC9041972.

One in twenty individuals worldwide suffer from depression,[1,2], and limited developments have been made in pharmacological treatments over the last four decades [3]. Current first-line treatment recommendations for major depressive disorder (MDD) involve medications that inhibit the reuptake of serotonin, norepinephrine, and dopamine through various mechanisms.[4] However, as indicated in the STAR*D study, roughly one in three patients fail to achieve clinical remission through these medications [5].It is known that a sufficient clinical response to these medications can take an upwards of 8 to 12 weeks [6]. Moreover, up to 15% of patients with MDD have a treatment-resistant form of the disorder [7]. Altogether, this highlights the urgent need for rapid-acting antidepressants with a novel mechanism of action.

It has recently been shown that repeated infusions of ketamine have rapid, cumulative, and sustained antidepressant effects [8]. It has also been shown that ketamine infusions can reduce suicidal ideation in treatment-resistant depression [9]. This antidepressant effect persists in racemic formulations, such as esketamine,[10] which shows non-inferiority to ketamine [11]. However, the exact mechanism underlying its rapid antidepressant and anti-suicidal effects remains unknown.

There is growing evidence that dysregulations in the glutamatergic and GABAergic systems are implicated in the pathophysiology of depression [12], which provides an opportunity for novel drug design and the repurposing of existing drugs. Ketamine has been shown to modulate extrasynaptic $GABA_A$ receptors in cortical neurons [13], and the rapid increase in glutamate that ketamine produces appears to be an essential component of its antidepressant effect [14].

While many candidate pathways have been proposed to mediate the antidepressant effects of ketamine,[15,16] few clinical trials have investigated biological predictors of treatment response. Among them, acute alterations in glutamate and glutamine levels, measured using in vivo

magnetic resonance spectroscopy, appears to mediate the antidepressant effects of ketamine [17]. However, no studies have yet identified a set of candidate biological markers that can predict treatment response to ketamine on an individual level. Clearly defined clinical markers in treatment-resistant depression coupled with effective, innovative, and fast acting treatments such as intranasal esketamine marks the dawn of precision psychiatry [18].

Acknowledgements

FDG has received grants from Fundacão de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG; APQ-02572-16 and APQ-04347-17), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 313944/2018-0), Emenda Parlamentar Federal (23970012), and Secretaria Nacional de Política sobre Drogas (01/2017). LCQ has received grants from Programa de Pesquisa para o SUS (CNPq/PPSUS/BA; 003/2017). FK has received grants from the Stanley Medical Research Institute (07TGF/1148), Instituto Nacional de Ciência e Tecnologia – Conselho Nacional de Desenvolvimento Científico e Tecnológico (INCT-CNPq; 465458/2014-9), and the Canadian Foundation for Innovation (CFI).

References

1. Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980 – 2013. Int J Epidemiol. 2014;43:476-93.
2. Merikangas KR, Jin R, He JP, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch Gen Psychiatry. 2011;68:241-51.
3. Hyman SE. Psychiatric drug development: diagnosing a crisis. Cerebrum. 2013;2013:5.
4. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 3. Pharmacological treatments. Can J Psychiatry. 2016;61:540-60.
5. Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps. Am J Psychiatry. 2006;163:1905-17.
6. Gelenberg AJ, Chesen CL. How fast are antidepressants? J Clin Psychiatry. 2000;61:712-21.
7. Berlim MT, Turecki G. Definition, assessment, and staging of treatment-resistant refractory major depression: a review of current concepts and methods. Can J Psychiatry. 2007;52:46-54.
8. Phillips JL, Norris S, Talbot J, Hatchard T, Ortiz A, Birmingham M, et al. Single and repeated ketamine infusions for reduction of suicidal ideation in treatment-resistant depression. Neuropsychopharmacology. 2020;45:606-12.
9. Phillips JL, Norris S, Talbot J, Birmingham M, Hatchard T, Ortiz A, et al. Single, repeated, and maintenance ketamine infusions for treatment-resistant depression: a randomized controlled trial. Am J Psychiatry. 2019;176:401-9.
10. Daly EJ, Singh JB, Fedgchin M, Cooper K, Lim P, Shelton RC, et al. Efficacy and safety of intranasal esketamine adjunctive to oral antidepressant therapy in treatment-resistant depression a randomized clinical trial. JAMA Psychiatry. 2018;75:139-48.
11. Correia-Melo FS, Leal GC, Vieira F, Jesus-Nunes AP, Mello RP, Magnavita G, et al. Efficacy and safety of adjunctive therapy using esketamine or racemic ketamine for adult treatment-resistant depression: a randomized, double-blind, non-inferiority study. J Affect Disord. 2020;264:527-34.
12. Duman RS, Sanacora G, Krystal JH. Review altered connectivity in depression: GABA and glutamate neurotransmitter deficits and reversal by novel treatments. Neuron. 2019;102:75-90.
13. Wang DS, Penna A, Orser BA. Ketamine increases the function of γ -aminobutyric acid type a receptors in hippocampal and cortical neurons. Anesthesiology. 2017;126:666-77.
14. Krystal JH, Sanacora G, Duman RS. Rapid-acting glutamatergic antidepressants – the path to ketamine and beyond. Biol Psychiatry. 2013;73:1133-41.
15. Krystal JH, Abdallah CG, Sanacora G, Charney DS, Duman RS. Ketamine: a paradigm shift for depression research and treatment. Neuron. 2019;101:774-8.
16. Zanos P, Gould TD. Mechanisms of ketamine action as an antidepressant. Mol Psychiatry. 2018;23:801-11.
17. Milak MS, Rashid R, Dong Z, Kegeles LS, Grunebaum MF, Ogden RT, et al. Assessment of relationship of ketamine dose with magnetic resonance spectroscopy of Glx and GABA

responses in adults with major depression a randomized clinical trial. JAMA Netw Open. 2000;3:e2013211.
18. Passos IC, Ballester P, Rabelo-da-Ponte FD, Kapczinski F. Precision psychiatry: the future is now. Can J Psychiatry. 2021 Mar 24;706743721998044. doi: http://10.1177/0706743721998044 Online ahead of print. » http://10.1177/0706743721998044

Chapter 6 - Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis

Authors: Devon Watts MSc[1]; Rafaela Fernandes Pulice[3]; Jim Reilly M.Eng., Ph.D[4]; Andre R Brunoni, PhD, MD[5]; Flávio Kapczinski, MSc, MD, PhD, FRCPC[1,3,6,7]; Ives Cavalcante Passos, MD, PhD[2,3]

1. Neuroscience Graduate Program, McMaster University, Hamilton, Canada

2- Universidade Federal do Rio Grande do Sul, School of Medicine, Porto Alegre, RS, Brasil.

3- Laboratório de Molecular Psychiatry, Centro de Pesquisa Experimental (CPE) and Centro de Pesquisa Clínica (CPC), Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, RS, Brasil.

4. Department of Electrical & Computer Engineering, McMaster University, Hamilton, ON, Canada.

5. Service of Interdisciplinary Neuromodulation, Laboratory of Neurosciences (LIM-27), Institute of Psychiatry, University of São Paulo, São Paulo, Brasil; Departamento de Clínica Médica, Faculdade de Medicina da USP, São Paulo, Brasil.

6. Instituto Nacional de Ciência e Tecnologia Translacional em Medicina (INCT-TM), Porto Alegre, RS, Brasil.

7. Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada.

*Corresponding author:
Ives Cavalcante Passos, MD, PhD
Professor of Psychiatry
Federal University of Rio Grande do Sul, Avenida Ramiro Barcelos, 2350, Zip Code: 90035-903, Porto Alegre-RS, Brasil, Phone: +55 512 101 8845, Email: ivescp1@gmail.com

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

Email for all authors:
Devon Watts: wattsd@mcmaster.ca
Rafaela Fernandes Pulice: rafaelafpulice@gmail.com
Jim Reilly: reillyj@mcmaster.ca
Andre Brunoni: brunoni@usp.br
Flávio Kapczinski: flavio.kapczinski@gmail.com
Ives Cavalcante Passos: ivescp1@gmail.com

This chapter in its entirety has been ***published*** in the journal **Translational Psychiatry**. The final accepted manuscript version of this article is presented within this thesis.

Watts, D., Pulice, R.F., Reilly, J. *et al.* Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Transl Psychiatry* **12**, 332 (2022). https://doi.org/10.1038/s41398-022-02064-z

ABSTRACT

Background: Selecting a course of treatment in psychiatry remains a trial-and-error process, and this long-standing clinical challenge has prompted an increased focus on predictive models of treatment response using machine learning techniques. Electroencephalography (EEG) represents a cost-effective and scalable potential measure to predict treatment response in Major Depressive Disorder.

Method: We performed separate meta-analyses to determine the ability of models to distinguish between responders and non-responders using EEG across treatments, as well as a performed subgroup analysis of response to transcranial magnetic stimulation (rTMS), and antidepressants (Registration Number: CRD42021257477) in Major Depressive Disorder by searching PubMed, Scopus, and Web of Science for articles published between January 1960 and February 2022.

Results: We included 15 studies that predicted treatment response among patients with major depressive disorder using machine-learning techniques. Within a random-effects model with a restricted maximum likelihood estimator comprising 758 patients, the pooled accuracy across studies was 83.93% (95% CI: 78.90-89.29), with an Area-Under-the-Curve (AUC) of 0.850 (95% CI: 0.747-0.890), and partial AUC of 0.779. The average sensitivity and specificity across models was 77.96% (95% CI: 60.05-88.70), and 84.60% (95% CI: 67.89-92.39), respectively. In a subgroup analysis, greater performance was observed in predicting response to rTMS (Pooled accuracy: 85.70% (95% CI: 77.45-94.83), Area-Under-the-Curve (AUC): 0.928, partial AUC: 0.844), relative to antidepressants (Pooled accuracy: 81.41% (95% CI: 77.45-94.83, AUC: 0.895, pAUC: 0.821). Furthermore, across all meta-analyses, the specificity (true negatives) of EEG models was greater than the sensitivity (true positives), suggesting that EEG models thus far better identify non-responders than to responders to treatment in MDD. Studies varied widely in

important features across models, although relevant features included absolute and relative power in frontal and temporal electrodes, measures of connectivity, and asymmetry across hemispheres.

Conclusions: Predictive models of treatment response using EEG hold promise in major depressive disorder, although there is a need for prospective model validation in independent datasets, and a greater emphasis on replicating physiological markers. Crucially, standardisation in cut-off values and clinical scales for defining clinical response and non-response will aid in the reproducibility of findings and clinical utility of predictive models. Furthermore, several models thus far have used data from open-label trials with small sample sizes and evaluated performance in the absence of training and testing sets, which increases the risk of statistical overfitting. Large consortium studies are required to establish predictive signatures of treatment response using EEG, and better elucidate the replicability of specific markers. Additionally, it is speculated that greater performance was observed in rTMS models, since EEG is assessing neural networks more likely to be directly targeted by rTMS, comprising electrical activity primarily near the surface of the cortex. Prospectively, there is a need for models that examine the comparative effectiveness of multiple treatments across the same patients. However, this will require a thoughtful consideration towards cumulative treatment effects, and whether washout periods between treatments should be utilised. Regardless, longitudinal cross-over trials comparing multiple treatments across the same group of patients will be an important prerequisite step to both facilitate precision psychiatry and identify generalizable physiological predictors of response between and across treatment options.

INTRODUCTION

It has been notably demonstrated in the Sequential Treatment Alternatives to Relieve Depression (STAR*D) study that antidepressants fail to facilitate remission in most patients with major depressive disorder (MDD), and that there is no clearly preferred medication when patients inadequately respond to several courses of antidepressants [1]. Similarly, data from a multicentre randomised controlled trial spanning 2439 patients across 73 general practices in the United Kingdom found that 55% of patients (95% CI: 53-58%) met the threshold for treatment resistant depression, defined as ≥14 on the BDI-II, and who had been taking antidepressant medication of an adequate dose, for at least 6 weeks [2].

This long-standing clinical challenge of selecting an appropriate treatment for any given patient has prompted the increasing development of predictive models of treatment response using machine learning techniques. Broadly speaking, supervised machine learning models use labelled training data (e.g., features or input variables), to predict a given outcome (e.g., treatment response) in unseen data (e.g., testing or validation dataset) [3]. In the context of psychiatry, these models have largely involved classification and regression tasks, where the outcome is a category (e.g., responders vs. non-responders), or a continuous outcome (e.g., depression change scores). There are several available algorithms to select from, each relying on a series of assumptions of the underlying input data. Moreover, an important consideration in model development is hyperparameter tuning, which involves finding a configuration of tuning parameters prior to model training that results in the best performance (e.g., accuracy for classification models, and lowest root mean squared error for regression models, respectively). A detailed overview of supervised machine learning [4], algorithm selection [3], and hyperparameter tuning [5]can be found elsewhere.

Thus far, most studies have utilised baseline clinical data to predict prospective treatment response at an individual level, with varying degrees of success and methodological robustness [6]. Similarly, there is a growing interest in the use of neuroimaging and neurophysiological markers as input features to these models. For instance, in a recent meta-analysis using MRI to predict treatment response in MDD, comprising 957 patients, the overall area under the bivariate summary receiver operating curve (AUC) was 0.84, with no significant difference in performance between treatments or MRI machines [7]. AUC, as described elsewhere, is a measure ranging from 0-1 indicating how well a parameter can distinguish between two diagnostic groups (e.g., responders/non-responders to an intervention).

However, fMRI and MRI remain impractical as widespread clinical tools to predict treatment response in psychiatry, considering high costs associated with each scan, and the excessive wait times to access a limited number of MRI machines. It was also recently shown in a landmark study that due to considerable analytical flexibility in fMRI pipelines, seventy independent teams yielded notably different conclusions when presented with the same dataset and series of hypotheses [8].

In contrast, measures such as electroencephalography (EEG) are comparably more cost-effective and scalable as a potential clinical tool to predict treatment response. As described elsewhere [9], EEG oscillations refer to rhythmic electrical activity in the brain and constitute a mechanism where the brain can regulate changes within selected neuronal networks. This repetitive brain activity emerges because of the interactions of large populations of neurons. As such, there is evidence that MDD may be related to abnormalities in large-scale cortical and subcortical systems distributed across frontal, temporal, parietal, and occipital regions [9].

For instance, power amplitudes in specific frequency bands, known as band power, are associated with different mechanisms in the brain. Although incompletely understood, alpha band power (8-12 Hz) reflects sensory and attentional inhibition and has been shown to be associated with creative ideation [10], beta frequencies (13-30 Hz) are prominent during problem solving [11], while delta frequencies (≤4 Hz) are notable during deep sleep [12], gamma frequencies (30-80 Hz) during intensive concentration [13], and greater theta band frequencies (4-8 Hz) during relaxation, respectively [14]. Alpha asymmetry, which measures the relative alpha band power between hemispheres, particularly within frontal electrodes, have been shown to discriminate individuals with MDD from healthy controls, although inconsistencies have been found across literature [15]. Similarly, beta and low gamma powers in fronto-central regions have been shown to be negatively correlated with inattention scores in MDD [16]. Moreover, intrinsic local beta oscillations in the subgenual cingulate were found to be inversely related to depressive symptoms, particularly in the lower beta range of ~13-25 Hz [17]. Additionally, in specific contexts, gamma rhythms, which represent neural oscillations between 25 and 140 Hz, have been shown to distinguish patients with MDD from healthy controls, and various therapeutic agents for depression have also been shown to alter gamma oscillations [18]. Patients with depression also show more random network structure, and differences in signal complexity [15], which may serve as replicable biomarkers of treatment response and remission.

A detailed description of potential EEG biomarkers of depression including signal features, evoked potentials, and transitions in resting-state EEG between wake and deep sleep, can be found elsewhere [15]. Altogether, no robust individual biomarker of treatment response in MDD has emerged. Towards this end, in a meta-analysis of treatment response prediction during a depressive episode, it was shown that the sensitivity across articles was 0.72 (95% CI=0.67-

0.76), and specificity was 0.68 (95% CI=0.63-0.73), respectively [19]. Nonetheless, most included studies used linear discriminant analysis in the absence of adequate cross-validation methods, training, and testing sets, or hyperparameter tuning, which may have led to biased performance metrics and a greater likelihood of statistical overfitting. Therefore, in the present study, we aimed to meta-analyse and systematically review studies that used machine learning techniques to predict treatment response in MDD.

METHODS

This study has been registered on PROSPERO with the registration number PROSPERO CRD42021257477.

Search strategy

Three electronic databases (PubMed, Scopus, and Web of Science) were examined for articles published between January 1960 and February 2022. To identify relevant studies, the following structure for the search terms was used: (Supervised Machine Learning OR Artificial Intelligence) AND (Major Depressive Disorder) AND (Electroencephalography) AND (Interventions OR Trials). The complete filter is available in the supplementary material. We also screened references from the included articles to identify potential missed articles. There were no language restrictions.

Eligibility criteria

This meta-analysis was performed according to the PRISMA statement [20]. We selected original articles that assessed patients with a psychiatric disorder treated with pharmacological or non-pharmacological interventions coupled with machine learning models and electroencephalography (EEG) features to predict treatment outcomes. Review articles and preclinical trials were excluded. A minimum criterion of cross-validation or training and testing

sets were required for study inclusion, since models lacking resampling procedures are less likely to appropriately generalise to independent datasets. Furthermore, studies with small sample sizes (≤30) that did not correct for overfitting were excluded, since cross-validation with small sample sizes, in the absence of training and testing sets, can lead to inflated and highly variable predictive accuracy [21]. Details relating to excluded studies can be found in Supplementary Table 1.

Data collection and extraction

Initially, the potential articles were independently screened for title and abstract contents by two researchers (DW and RFP). Then, they also obtained and read the full text of potential articles. A third author (ICP) provided a final decision in cases of disagreement. Data extracted from the studies included publication year, sample size, diagnosis, EEG system, reference choice, impedance, number and type of electrodes, method for de-artificing, feature selection and extraction method, type of intervention, outcomes of interest, machine learning algorithm, and performance metrics of the models (i.e., accuracy, balanced accuracy, sensitivity, specificity, area under the curve, true positive, false positive, true negative and false negative, and coefficient of determination). We also developed a quality assessment instrument specific to machine learning studies since there is no tool for quality assessment in machine learning studies. Briefly, the quality assessment evaluates studies according to several domains including representativeness of the sample, confounding variables, outcome assessment, machine learning approach, feature selection, class imbalance, missing data, performance/accuracy, and testing/validation. This instrument, and a brief description of each component, are further described in the Supplementary Material. Additionally, we utilised the Quality Assessment of

Diagnostic Accuracy Studies-2 (QUADAS-2) [22] to assess potential bias and variation in each included study, as described in Supplementary Table 2.

In terms of the analysis, "mada" [23], "dmetatools" and "meta" packages in R were used to meta-analyse diagnostic accuracy studies. The metamean function in the "meta" package was used to pool accuracy across studies in a random effects model using an inverse variance method with Knapp-Hartung adjustments to calculate the confidence interval around the pooled effect. A restricted maximum-likelihood estimator was used to calculate the heterogeneity variance τ^2. Moreover, the madad function in the "mada" package was used to calculate the sensitivity, specificity, and pAUC across studies, while the madauni function was used to calculate the Diagnostic Odds Ratio (DOR), positive likelihood ratio (posLR), and negative likelihood ratio (negLR). AUC was calculated using the AUC_boot function in dmetatools, with an alpha of 0.95 and 2000 bootstrap iterations.

RESULTS

We found 2489 potential abstracts and included 15 articles in the present meta-analysis and systematic review, two included after reference screening (Supplementary Table). A list of included studies as well as their most relevant characteristics and findings are detailed in Table 1. Two separate quality assessments can be observed in the supplementary material. Of the included studies, seven predicted response to brain stimulation therapies [24–30], and eight predicted response to pharmacological treatment [31–38]. Additionally, a complete breakdown of how each study defined treatment response can be found in Supplementary Table S4.

Studies predicting treatment response to brain stimulation therapies

There were seven studies using EEG features to predict treatment response to brain stimulation [24–30]. Among these, all predicted response to repetitive transcranial magnetic stimulation (rTMS).

Further information relating to feature extraction methods, feature selection, and extracted features can be found in Table 2.

Corlier and colleagues predicted treatment response to open label 10 Hz rTMS applied to the left dorsolateral prefrontal cortex (DLPFC) in a sample of 109 patients with MDD. Treatment response was defined as a decrease of ≥40% in post-treatment 30-item inventory of depressive symptomatology—self-rated (IDS-30) scores. Extracted features comprised changes in neurophysiological connectivity in the individual alpha frequency (IAF) band in response to rTMS stimulation. Using an elastic net model, which provides an embedded form of feature selection, the authors reported an accuracy of 61.8-69.3%, with the best performance using alpha spectral coherence features, defined as spectral correlation in the alpha frequency band. Of note, the same model showed 77% accuracy in a unilateral treatment subgroup [26].

Furthermore, Erguzel and colleagues developed a model to predict antidepressant response to 20 sessions of adjunctive 25 Hz rTMS applied to the left PFC in a sample of 147 individuals with MDD. Responder status was operationalized as a ≥50% reduction in Hamilton Depression Rating Scale (HAM-D) scores at the end of treatment. The best performance was observed in a Support Vector Machine (SVM) model in the theta frequency band across prefrontal regions using cordance features, which combines absolute and relative resting EEG activity, with an accuracy of 86.4% [29]. Additionally, Hasanzadeh et al. developed a model to predict response to 5-sessions of 10 Hz rTMS applied to the left DLPFC among 46 patients with MDD. Treatment response was defined as ≥50% decrease in BDI-II or HAMD-24 scores, or by BDI ≤ 8 (HAMD-24 ≤ 9) which indicates remission. Using a k-Nearest Neighbours (k-NN) model, the best performance was observed using Lempel-Ziv complexity features in the beta frequency band, which counts the number of distinct segments in the signal, with an accuracy of 82.6%.[30].

Another study [27] predicted treatment response (≥50% improvement in HAMD-17) in an 18-session open-label trial of 25 Hz rTMS to the left prefrontal cortex, comprising 55 patients with MDD using cordance features in the delta and theta frequency bands, resulting in 89.09% accuracy. However, since accuracy was assessed using internal k-fold cross-validation alone, performance may be over-optimistic. In another study, treatment response was predicted within a 15-session open-label trial of 10 Hz left prefrontal rTMS in 39 patients with MDD using theta, upper alpha, and upper gamma power and connectivity, as well as theta-gamma coupling features, resulting in an accuracy of 91% [24]. Similarly, in another study using the same experimental design in 32 patients with MDD, treatment response was predicted using theta and alpha power and connectivity, frontal theta cordance, and alpha peak frequency, resulting in an accuracy of 86.66% [25]. Furthermore, other studies with insufficient sample sizes predicted response to tDCS [39], and rTMS [40], as further described in Supplementary Table S1.

Across neurostimulation trials, important features included absolute and relative power in frontal electrodes (alpha and theta band), connectivity measures (theta and gamma), spectral entropy, and cordance features across alpha, theta, delta, and gamma frequency bands. As described elsewhere [41], spectral entropy of a signal is a measure of its spectral power distribution and is based on Shannon's entropy. With respect to important channels, one study [27] found Fp1, Fp2, F3, F7, and F8 in the theta frequency band to be important features following feature selection, and these same features were used in a follow-up study [29] by the same group, largely maintaining model accuracy (89.12% vs 78.3-86.4%, respectively). One study [30] compared nonlinear, power spectral density, bi-spectral features, and cordance, with the best performance observed when restricting features to power over all 19-channels in delta, theta, alpha and beta frequency ranges. Furthermore, another study [24] found enhanced theta power at Fz to be significantly

different between responders and non-responders (F1=8.577, p=0.006), however no main effect for frontal-midline theta power was observed in a follow-up study [25]. Furthermore, three studies [24,25,29] did not report feature selection methods, and surprisingly, no studies compared multiple feature selection methods. Further details can be observed in Table 2.

Studies predicting clinical response to pharmacological treatment

Seven studies developed predictive models of clinical response to pharmacological treatment [31–38]. Among these, three studies assessed treatment response to various classes of antidepressants within randomised double-blind trials [32–34,37], one assessed response within a randomised trial of ketamine or placebo [31], one assessed response in an open-label trial of an SSRI [42], and two other studies assessed response to sertraline [37], and escitalopram [38], respectively.

Wu and colleagues developed a machine learning model known as Sparse EEG Latent SpacE Regression (SELSER), applied to alpha, beta, delta, and gamma frequency bands, to predict antidepressant treatment response using resting state EEG. SELSER was first trained on data from the largest neuroimaging-couped placebo-controlled randomised clinical study of antidepressant efficacy, comprising 309 patients. The generalizability of the antidepressant signature was tested in two independent samples of depressed patients treated with antidepressants, and another sample of patients treated with rTMS to assess the specificity of SELSER's signature for predicting response to antidepressants. Response was defined according to HAMD-17 change scores at the end of treatment. SELSER was shown to generalise across antidepressant datasets, with an R2 of 0.60 in predicting response to sertraline, and an R2 of 0.41 in predicting response to placebo, respectively [37].

Cao and colleagues developed a machine learning model to predict rapid antidepressant response to ketamine in a sample of 55 patients with treatment resistant depression. Response was defined as ≥ 45% reduction in depressive symptoms (HAMD-17) 240 minutes following infusion. Using EEG power in delta, theta, lower alpha, and upper alpha bands, as well as alpha asymmetry in frontal electrodes as candidate features, the best performance was observed using SVM with a radial kernel, resulting in an accuracy of 78.4% [31].

De la Salle and colleagues developed a model to predict response within a double-blinded 12-week trial of escitalopram, bupropion, or combined treatments, in 47 patients with treatment resistant depression. Clinical response was defined as a ≥ 50% reduction in MADRS scores from baseline, and remitters were operationalized as those with ≤ 10 MADRS scores at posttreatment. Within a logistic regression model, change scores in middle right frontal cordance and prefrontal cordance across delta, theta, alpha, and beta frequency bands resulted in an accuracy of 74% and 81% in predicting clinical response, respectively. Similarly, clinical remission could be predicted with 70% accuracy using prefrontal cordance, however middle right frontal cordance features were not discriminative (51% accuracy). It is important to note that EEG features alone resulted in better accuracy (74-81%) than clinical features alone (66%) or a combined model of EEG and clinical features (64-66%) [33].

Furthermore, Zhdanov et al. predicted antidepressant response to an 8-week open-label trial of escitalopram (10-20 mg) in a sample of 122 patients with MDD. Patients were classified as responders if they showed ≥50% reduction in Montgomery-Asberg Depression Rating Scale (MADRS) scores at the end of treatment. Of note, four classes of features were used, comprising electrode-level and source-level spectral features, multiscale-entropy-based features, and microstate-based features, as described in further detail within Supplementary Table 1. Using

baseline EEG features alone, their SVM model showed an accuracy of 79.2%. Performance improved slightly when adding EEG features from the second week of treatment, with an accuracy of 82.4% [38].

In another study, Rajpurkar and colleagues predicted improvement in individual symptoms within the HAM-D from baseline to week 8 within a randomised trial of escitalopram, sertraline, or extended-release venlafaxine in a sample of 518 patients with MDD. Pre-treatment EEG candidate features included frontal alpha asymmetry, occipital beta asymmetry, and the ratio of beta/alpha and theta/alpha band power for each electrode. Using a gradient boosting machine (GBM) model with embedded feature selection, the authors reported an R2 of 0.375-0.551, with the best performance using EEG and baseline symptom features [36]. Other studies predicted response to various classes of antidepressants, resulting in an accuracy of 88% [34], treatment remission, resulting in an accuracy of 64.4% [32], and treatment response to an open-label trial of an SSRI, resulting in an accuracy of 87.5% [35].

Across medication trials, important features included alpha, theta, and gamma power in frontal electrodes, coherence between frontal and temporal electrodes, change scores in delta power, ratio of alpha and theta power in temporal electrodes, and asymmetry between hemispheres. With respect to important channels, two studies [31,36] found Fp2 absolute theta to be among the top ten features to predict response to SSRIs/SNRI, and ketamine, respectively. Additionally, two studies [34,36] showed baseline power at F7 to be an important feature, although in different frequency bands, corresponding to alpha, and beta and gamma, respectively. Overall, studies varied widely in the number of electrodes, electrodes of interest, and feature extraction methods, which preclude a set of well-elucidated individual biomarkers of treatment response.

Improvements in model accuracy by incorporating EEG features

Additionally, we sought to investigate the contribution of EEG-based features to predictive accuracy in cases where clinical variables were also incorporated into predictive models of treatment response. However, only six studies [24–26,34,36,38] (42.8%) used both EEG and clinical candidate features within model development. Among them, only one [26] reported differences in model accuracy between EEG features, clinical features, and combined models. Corlier and colleagues reported that alpha spectral correlation features predicted treatment response with 69.3% accuracy (Sensitivity: 67.1%, Specificity: 70.9%), while baseline IDS-30 scores predicted treatment response with 75.1% accuracy (Sensitivity: 64.1%, Specificity: 83.6%). Combining both features lead to greater model performance, with an accuracy of 79.2% (Sensitivity: 75.7%, Specificity: 81.9%) [26].

Quality Metrics

Overall, samples used to develop models were small, with a median sample size of 55 among studies predicting response to neurostimulation, and 86.5 among studies predicting response to antidepressant medication, respectively. Quality metrics were assessed using the QUADAS-2 [22], and a quality assessment instrument specific to machine learning. These quality assessment metrics can be found in Supplementary Table 2, and the Supplementary Material, respectively. The QUADAS-2, as described elsewhere[22], evaluates risk of bias according to the domains of patient selection, index test, reference standard, and flow and timing. Overall, most studies showed low risk of bias according to patient selection, how treatment response was defined, and the time interval between EEG assessments and treatment follow-up. However, 7 of 15 (46.6%) [24–26,29,30,33] showed a high risk of bias in reference standards for model development, which

included a lack of training/testing sets, and lack of blinded assessment to treatment allocation when collecting symptom scales and EEG data.

With respect to the machine learning quality assessment, the median score for neurostimulation studies was 5/9 (55.5%), and the median score for psychiatric medication studies was 6.5/9 (72.2%), respectively. Most studies [24–33,35,38] did not discuss methods to address class imbalance, which occurs in classification models where there is a disproportionate ratio of observations in each class (e.g., responders vs non-responders). Moreover, several studies [24,25,27–30,32–34,36] evaluated performance using cross-validation in the absence of training and testing sets, which increases the risk of model overfitting, and may lead to biased results.

Meta-analyses of predictive models of treatment response using EEG

Within the fifteen studies included in the systematic review, seven predicted treatment response to rTMS [24–30], and eight predicted response to antidepressant treatments (ketamine, escitalopram, sertraline, escitalopram, bupropion, and venlafaxine) respectively [31–34,36–38,43]. Among them, eleven involved binary classification models [24–26,28–30,32,33,44–46](response vs non-response) and reported summary statistics required to pool predictive accuracy. A detailed summary of performance metrics across models can be found in Supplementary Figure S4. The accuracy of treatment response prediction models in MDD across 758 patients was pooled in a random-effects model using an inverse variance method with restricted maximum likelihood estimator to calculate the heterogeneity variance τ^2. Furthermore, Knapp-Hartung adjustments were used to calculate the confidence interval around the pooled effect.

Overall, across six studies comprising 438 patients with MDD, the pooled accuracy of treatment response prediction using EEG was 83.93% (95% CI: 78.90-89.29), with a heterogeneity variance τ^2 of 0.0044 (95% CI: 0.0009-0.0296), as depicted in Figure 1. Moreover, the median

sensitivity across studies was 77.96% (95% CI: 60.05-88.70), and median specificity was 84.60% (95% CI: 67.89-92.39), respectively. Additionally, as shown in Table 3, the AUC was 0.850 (95% CI: 0.747-0.890), with a pAUC of 0.777, whereas the total DOR was 23.49 (95% CI: 10.40-52.02), with a posLR of 5.232 (95% CI: 3.15-8.67), and negLR of 0.271 (95% CI: 0.195-0.376), respectively. Briefly, DOR is a ratio of the odds of testing positive (e.g., predicted as a responder) when actually reaching therapeutic response to treatment, relative to the odds of testing positive (e.g., predicted as a responder), when failing to respond to treatment, although this metric is also dependent on prevalence [47]. Further information regarding this metric can be found elsewhere [48]. Similarly, posLR describes the probability of testing positive divided by the probability a positive test would be expected in a negative case, whereas negLR is defined as the opposite. A posLR of 10 or more and a negLR of 0.1 or less are generally deemed to be informative tests. Additionally, considering potential study heterogeneity across treatment modalities, a subgroup analysis was performed for rTMS and antidepressant models, where these outcomes were assessed separately, as shown in Supplementary Figures S1-S4.

Efficacy of predicting treatment response to rTMS

Across six studies [24–26,28–30], comprising 438 patients, the pooled accuracy of rTMS treatment response prediction using EEG was 85.70% (95% CI: 77.45-94.83), with a heterogeneity variance τ^2 of 0.0051 (95% CI: 0.0004: 0.0668). The median sensitivity across studies was 79.4% (95% CI: 58.65-90.80) and median specificity was 92.05% (95% CI: 81.70-99.30), respectively. Overall, the AUC across studies was 0.895 (95% CI: 76.07-93.99), with a partial AUC of 0.821, a DOR of 35.48 (95% CI: 7.805-161.364, τ^2=2.797), posLR of 7.098 (95% CI: 2.843-17.725, τ^2=0.915), and negLR of 0.234 (95% CI: 0.122-0.448, τ^2=0.478), respectively.

A test for equality of proportions with a continuity correction of 0.5 yielded a Chi-Squared (X^2) value of 20.05 (p= 0.0012) and 20.62 (p= 0.00095) for sensitivities and specificity, respectively. Moreover, a moderate negative correlation was observed between sensitivities and false positive rates (Rho = -0.526 (95% CI: -0.937 – 0.498). Further details can be observed in Supplementary Figures S1 and S3.

Efficacy of predicting treatment response to antidepressants

Across five studies, comprising 325 patients, the pooled accuracy of antidepressant treatment response prediction using EEG was 81.41% (95% CI: 71.09-92.23), with a heterogeneity variance τ^2 of 0.0052 (95% CI: 0.00-0.11), as depicted in Supplementary Figure S2. The median sensitivity across studies was 77.78% (95% CI: 61.14-88.50), and median specificity was 82.06% (95% CI: 65.54-95.24), respectively. Overall, the AUC of studies predicting response to antidepressant medications was 0.764 (95% CI: 0.710-0.899) with a partial AUC of 0.756. Furthermore, the overall DOR was 19.02 (95% CI: 5.51-65.61), with a posLR of 4.30 (95% CI: 1.92-9.64), and negLR of 0.296 (95% CI: 0.208-0.422). A test for equality of proportions with a continuity correction of 0.5 yielded an X^2 of 3.8 (p=0.434) for sensitivities and an X^2 of 23.67 (p=0.0000927) for specificities, respectively. Moreover, a weak negative correlation of sensitivities and false positive rates were observed across studies (Rho = -0.016, 95% CI: -0.886-0.879). Further details can be observed in Supplementary Figures S2 and S4.

Considering the small number of antidepressant studies, we performed another meta-analysis with the addition of three studies [49–51] that were excluded due to a small sample size ($N \leq 30$), increasing the total sample to 402 patients with MDD. This resulted in a pooled accuracy of 84.52% (95% CI: 77.67-91.98, τ^2= 0.0034), median sensitivity of 82.07% (95% CI: 60.96-91.72), median specificity of 84.47% (95% CI: 65.28-92.55), and AUC of 0.794 (95% CI: 0.728-

0.887). Additionally, the DOR was 28.98 (95% CI: 9.95-84.4), with a posLR of 5.20 (95% CI: 2.67-10.15), and negLR of 0.26 (95% CI: 0.19-0.37). Further details can be found in Supplementary Figure S5.

DISCUSSION

While there is a great deal of promise in using EEG within machine learning models to predict treatment response in MDD, there does not appear to be a consensus on collection methods, or consistent physiological markers of response to antidepressants, or rTMS across studies. Given the complexity of MDD, and the likelihood of heterogeneity in important features across patients, the field may require a conceptual shift away from the search for singular biomarkers, towards the use of composite features, identified using multivariate models. As such, it may be the case that no singular neurophysiological biomarker will demonstrate the sensitivity and specificity required to guide treatment selection in MDD. Rather, a composite biomarker comprising a series of distinct, but mutually informative features, may serve to both improve our mechanistic understanding of treatment response, and appropriately model this phenomenon. However, it is important to highlight that multimodal feature combinations carry several additional considerations. Namely, if complex approaches such as source localization are required to provide meaningful accuracy, this may provide a significant challenge in the clinical implementation of such models. Additionally, while resting-state features provide greater scalability relative to EEG activation patterns during specific tasks, the latter may inform features that could perhaps be more sensitive and specific in modelling clinical improvement in response to a given treatment.

Model performance across meta-analyses

Overall, model performance in predicting response to rTMS (accuracy = 85.70%, 95% CI: 77.45-94.83; AUC = 0.895, 95% CI: 76.07-93.99, DOR = 35.48, 95% CI: 7.805-161.364) was greater than predicting response to antidepressants (accuracy = 81.41%, 95% CI: 71.09-92.23; AUC = 0.764, 95% CI: 0.710-0.899, DOR = 19.02, 95% CI: 5.51-65.61), even after the addition of three excluded studies to increase the sample size (accuracy = 84.52%, 95% CI: 77.67-91.98; AUC = 0.794, 95% CI: 0.776-0.919; DOR = 28.98, 95% CI: 9.95-84.4). This was also found relative to a total model including 12 studies (N=792) across all rTMS and medication trials (accuracy = 83.93%, 95% CI: 78.90-89.29; AUC: 0.850, 95% CI: 0.600-0.887; DOR = 23.49, 95% CI: 10.40-52.02).

There are several potential contributing factors to this finding, as models that predicted response to rTMS utilised data from open-label trials that lacked an adequate sham condition. However, it is posited that this may be reflective of very specific targets across rTMS studies, since all involved high-frequency stimulation (10-25 Hz) to the DLPFC. Moreover, it is speculated that EEG, which measures electrical activity primarily near the surface of the cortex, is assessing neural networks that are more likely to be directly targeted by rTMS. Conversely, with respect to pharmacotherapy, the effect is much more indirect and potentially dependent on other factors that EEG cannot access such as hepatic metabolism, and pharmacokinetic interactions.

Interestingly, across all four meta-analyses, model specificity (82.06-92.05%) was notably greater than model sensitivity (77.96-82.07%), even when considering the upper and lower bounds of the confidence intervals. This suggests that across all treatment modalities, including rTMS, antidepressants, and a combined model, EEG features are better able to capture predictors of clinical non-response to treatment, rather than predictors of clinical response. As such, it is possible that EEG may show greater utility in determining whether a patient will not respond to a

given intervention at baseline. However, prospective validation with large samples in independent cohorts will be necessary to determine the reliability of this finding.

Additionally, the rTMS model showed a higher DOR (DOR=35.48, 95% CI: 7.805-161.364; $\tau 2$=2.797, 95% CI: 0.00-8.402), relative to the total model (DOR=23.49, 95% CI: 10.40-53.02; $\tau 2$=1.395, 95% CI: 0.00-2.13), and antidepressant model (DOR=19.02, 95% CI: 5.51-65.61); $\tau 2$=1.27, 95% CI: 0.00-14.79), respectively. This indicates that the odds for positivity among individuals who respond to treatment is 35 times higher than the odds for positivity among individuals who will not respond to treatment. However, it is important to highlight that a large upper and lower bound of the confidence interval was observed across rTMS studies, as well as greater heterogeneity.

Independent validation, feature replicability, and clinical outcomes

Nonetheless, there is a need for greater emphasis on testing model performance with independent samples, greater consistency in sample collection and model development, and an increased focus on replicating features identified in previous models. Additionally, nine studies [24–30,33] (60%) included in the present meta-analysis and systematic review did not test accuracy in holdout data, relying instead on internal cross-validation, which may lead to overoptimistic performance metrics. Furthermore, most studies (57.1%) utilised data from open-label trials lacking adequate double-blind procedures, and as such, there is a risk of bias pertaining to the scoring and interpretation of treatment response. There also remains an unmet need for prospective studies that compare features between models of treatment response and remission outcomes. Thus far, only one study [33] has assessed both outcomes, although it did not report a difference in top features between these models. It remains to be determined whether there are

reproducible features that are specific to reaching threshold for treatment response, relative to treatment remission.

Definitions of clinical response

A majority of studies contained in the present review (86.6%) used binary classification models to discriminate treatment responders' treatment from non-responders. As detailed further in Supplementary Figure S4, studies varied in terms of the specific clinical scale and change-score thresholds that constituted treatment response. Overall, four studies (26.6%) selected a ≥ 50% reduction on the HAMD-17 as the threshold of clinical response, while three studies (20%) defined clinical response as ≥ 50% reduction on the MADRS. Large differences in treatment duration were also observed across trials. Importantly, greater standardisation in how clinical response is defined is required to better assess the performance of prospective models, aid in the reproducibility of findings, and improve the likelihood of real-world clinical utility of ML models in psychiatry. Similarly, as described elsewhere [52], there is a lack of clear consensus on how treatment resistance is defined, which highlights the need for greater consistency across studies.

Comparison of algorithms across studies

Furthermore, only three studies (20%) [28,44,53] assessed the performance of multiple algorithms, which limits a comparison on which algorithms tended to perform well. Considering this, two studies [39,40] that were excluded due to insufficient sample size which assessed multiple algorithms were pooled with included studies to examine potential trends, comprising a total of five studies. Among them, SVM was compared alongside other algorithms such as random forest within five studies and resulted in the best performance in 60% of cases. In the other 40% of cases [39,53], only composite accuracy across algorithms was reported. As described elsewhere [54],

SVM is well suited to very high dimensional data, considering its use of support vectors, various available kernels, and computational efficiency in large datasets.

Pre-processing strategies across studies

With respect to pre-processing strategies, all studies used a bandpass filter to limit included frequencies to a specific range, although studies varied widely (0.1-80 Hz) in terms of the upper and lower bounds. One study [38] also reported using a notch filter at 60 Hz, which attenuates frequencies in a specific range to very low levels. Furthermore, five studies [27–29,36,53] (33.3%) used independent component analysis to filter artefacts, and five [27–29,53,55] (33.3%) used a fast Fourier transform method. Other studies [30,38] used available pre-processing packages, such as the EEGLAB toolbox available in the MATLAB programming language.

Future Perspectives

Prospectively, there is a need for models that examine the comparative effectiveness of multiple treatments across the same patients. Studies thus far have focused on predicting response to a specific intervention rather than treatment selection, and few have been replicated to see if a classification tool has worked in external independent datasets.

Furthermore, to facilitate EEG biomarkers of response to specific treatments, future studies may benefit from testing model performance on external datasets of other psychiatric medications or neurostimulation therapies. For example, Wu and colleagues assessed whether the algorithm SELSER, trained on SSRI datasets, could predict response to rTMS [37]. This approach may help highlight differences in important features to predict treatment response across psychiatric medications and provide an avenue to investigate potential neurophysiological mechanisms of action. Moreover, by exploring whether models retain similar features and modest prediction accuracy when tested on external datasets of other interventions, this may provide a way to

identify generalizable EEG biomarkers that are related to therapeutic improvement or treatment resistance across disorders. Nonetheless, it may be more informative and realistic to focus on predictors of response to specific classes of medications and neurostimulation trials, to identify divergent mechanisms of therapeutic efficacy and treatment resistance. Either way, this will require a careful consideration of differences in outcome instruments between datasets.

Surprisingly, in the present review there was little overlap in top features between models, even when stratifying between rTMS or antidepressant trials. As such, there remains a critical need for a systematic comparison of several types of features in prospective models of treatment response and treatment selection to help guide prospective biomarker identification and validation. Of the fifteen studies comprising the current review, only three [30,38,56] (20%) included three or more categories of candidate features during model development. For instance, Hasanzadeh and colleagues considered nonlinear, spectral entropy and cordance features, and found that combining spectral entropy (beta and delta) and cordance features resulted in the highest performance [30]. Furthermore, Zhdanov and colleagues compared electrode-level spectral features, source-level spectral features, multiscale-entropy-based features, and micro-state-based features. Here, multiple-entropy-based features comprised the top 4 of 8 features in a model to predict response to 8-weeks of open label escitalopram [38].

Apart from the categories of features used in the present review, as detailed in Table 2, prospective models may benefit from incorporating features derived from brain source localization methods. This process, as described elsewhere [57], involves predicting scalp potentials from current sources in the brain (forward problem) and estimating the location of the sources from measuring scalp potentials (inverse problem). These methods have the potential to improve the signal-to-noise ratio of extracted features and suppress volume conduction.

However, they require an accurate head model which is often difficult to obtain. It remains unclear what the overall effectiveness of these methods are in the context of extracting meaningful features to predict treatment response.

Furthermore, as described in Supplementary Table S5, most predictive models have been developed using features derived from resting-state EEG. Only two studies [24,35] (13.3%) have used task-specific EEG to derive features, which involved the Sternberg Working Memory Task and 3-Stimulus Visual Oddball Task. Apart from this, event-related potentials may prove useful, especially if we could identify stimuli that are sensitive to depressed and psychotic states. Moreover, none of the reviewed studies developed predictive models using a combination of resting state and task-specific EEG. Incorporating both within the same model of treatment response may help inform potential mechanisms of action and yield more informative biomarkers. Additionally, no studies thus far have utilised intracranial EEG to predict treatment response in MDD. By placing electrodes directly on the surface of the brain, intracranial EEG provides a much cleaner signal, and by its nature, greater source localization [58]. While intracranial EEG is much more invasive relative to surface electrodes, they may be justified for severe cases of treatment resistance.

With respect to algorithm selection, SVM was found to perform well when comparisons against other algorithms were available. Apart from the approach of comparing performance across individual algorithms, stacked generalisation [59] provides an alternative ensemble method to combine the predictions of two or more machine learning algorithms, while using another algorithm to learn how to combine their outputs. As described elsewhere [60], stacking can improve model performance over any single model contained in the ensemble. Additionally, stacking differs from the traditional bagging and boosting ensemble methods in that it typically

uses different models that combine predictions from contributing models, rather than a series of decision trees, or models that comprise weak learners building upon the prediction of previous models, respectively. While two studies [39,53] averaged results across models into a composite accuracy, to our knowledge, stacked generalisation has not yet been explored in predictive models of treatment response using EEG.

Similarly, hyperparameter tuning, which involves selecting the optimal set of hyperparameters for a given model, remains an important consideration in model development [61]. While many software packages have default hyperparameter settings during cross-validation, searching the hyper-parameter space for the lowest loss-function, or best cross-validation score, is recommended. Although an exhaustive search of the hyperparameter space is often computationally infeasible, there are several available methods such as a manual grid search, collaborative hyperparameter tuning [62], and Bayesian optimization [63].

As demonstrated in the current review, studies varied largely in the number of electrodes used, EEG systems, feature selection and extraction methods, and machine learning algorithms. Considering the heterogeneity observed across studies, large, standardised datasets must become available before this field can move ahead in a significant way. Importantly, there is a need for models developed using large well-characterised samples, with separate training, testing, and external validation datasets, to derive classification tools that can be useful clinically. Similarly, available repositories are needed to appropriately replicate models developed thus far, identify generalizable biomarkers of treatment response across interventions, and identify distinct neurophysiological markers that can help guide treatment selection in MDD.

Conflict of interest statements

Ph.D. Thesis – D. P. Watts; McMaster University – Neuroscience.

Devon Watts, Rafaela Fernandes Pulice, Jim Reilly, Andre R Brunoni, and Ives Cavalcante Passos report no biomedical financial interests or potential conflicts of interest. Flávio Kapczinski has received grants/research support from AstraZeneca, Eli Lilly, Janssen-Cilag, Servier, NARSAD, and the Stanley Medical Research Institute; has been a member of the speakers' boards of AstraZeneca, Eli Lilly, Janssen and Servier; and has served as a consultant for Servier.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Furthermore, this work received financial supports from Conselho Nacional de Desenvolvimento Científico e Tecnológico. We would also like to thank anonymous reviewers for their helpful feedback.

First author, year	Sample size and diagnosis [1,2]	Intervention	Outcome	Machine learning model	Accuracy	Other measures
STUDIES PREDICTING RESPONSE TO NEUROSTIMULATION THERAPY						
Bailey, 2017	39 patients with treatment-resistant depression	3 weeks (15 sessions) unilateral left 10 Hz rTMS	Responders vs. Non-responders Responders defined as ≥ 50% decrease in HAM-D after 5-8 weeks of rTMS	*Linear* SVM	91%	Sensitivity:91% Specificity:92% F1 score: 0.93
Bailey, 2018	32 patients with treatment-resistant depression	3 weeks (15 sessions) unilateral left 10 Hz rTMS	Responders vs. Non-responders Responders defined as ≥ 50% decrease in HAM-D after 5-8 weeks of rTMS	*Linear SVM*	86.66%	Sensitivity: 84% Specificity: 89%
Corlier, 2019	109 patients with MDD	3 weeks (15 sessions) of 10 Hz left DLPFC rTMS (68 subjects received unilateral left treatment, 41 were changed to sequential bilateral treatment – 10 Hz left DLPFC, 1 Hz right DLPFC)	Responders vs. Non-responders Responders defined as ≥ 40% decrease in IDS-30 scores from baseline to treatment 30	Elastic Net	61.8%-79.2% *(best performance observed with alpha band frequency and IDS-30 percent change score)*	AUC: 0.52-0.77 Specificity: 70.9-82.7% Sensitivity: 34.8-75.7% PPV: 58.2-79.7% NPV: 63.8-82.2%

Erguzel, 2014	147 patients with treatment-resistant depression	18 sessions of 25 Hz left PFC rTMS	Responders vs. Non-responders Responders defined as ≥ 50% decrease in HAM-D scores after 3 weeks of treatment	BPNN	89.12%	Sensitivity: 94.44% AUC: 0.904
Erguzel, 2015	55 patients with MDD	18 sessions of 25 Hz left PFC rTMS	Responders vs. Non-responders Responders defined as ≥ 50% decrease in HAM-D scores after 3 weeks of treatment	ANN	89.09%	Sensitivity: 86.67-93.33% Specificity: 80-84% AUC: 0.686-0.909 Best model (6-fold CV) Sensitivity: 93.3% Specificity: 84.0% AUC: 0.909
Erguzel, 2016	147 patients with treatment-resistant depression	20 sessions of adjunctive 25 Hz left PFC rTMS	Responders vs. Non-responders Responders defined as ≥50% decrease in HAM-D scores after 20 sessions of rTMS	ANN SVM DT	Accuracy: 78.3-86.4% *Best performance using SVM* Balanced Accuracy: 54.71-75.42%	Sensitivity: 60.41-68.62% Specificity: 49.01-82.22%
Hasanzadeh, 2019	46 patients with MDD	5-sessions of 10 Hz left DLPFC rTMS	Responders vs. Non-responders Remission vs. Non-remission Responders defined as ≥50% decrease in BDI-II or HAM-D	kNN	76.1-91.3% *best performance with power spectral features*	Sensitivity: 69.6-87% Specificity: 82.6-95.7%

			scores from baseline Remission defined as BDI ≤8 or HAM-D ≤9			
STUDIES PREDICTING RESPONSE TO PHARMACOLOGICAL TREATMENT						
Cao, 2019	37 patients with treatment-resistant depression	Patients randomised to one of three groups (1:1:1): 1. 0.5 mg/kg ketamine 2. 0.2mg/kg ketamine 3. Normal saline	Responders vs. Non-responders Responders defined as ≥45% reduction in HAM-D score from baseline to 240 min posttreatment	LDA NMSC kNN PARZEN PERLC DRBMC SVM *Radial kernel*	78.4% *Best performance using SVM with a radial kernel*	Sensitivity: 79.3% Specificity: 84.2% Recall: 78.5% Precision: 87.0% F1 score: 52.6%
Cook, 2020	180 patients with MDD	8-week trial of escitalopram (10mg) or bupropion (150mg) (1 week single-blind escitalopram followed by 7 weeks double-blind trial)	Remission vs Non-remission Remission defined as ≤7 HDRS at week 8	LDA	64.4%	Sensitivity: 74.3% Specificity: 55.3% PPV: 60.5% NPV: 70.0% AUC: 0.635
de la Salle, 2020	47 patients with MDD	12-week double-blinded trial of: 1) escitalopram + bupropion 2) escitalopram + placebo 3) bupropion + placebo	Responders vs. Non-responders Responders defined as ≥50% reduction in MADRS scores from baseline to posttreatment Remitters/Non-	LR	*Response*: Change in PF Cordance: 81% Change in MRF Cordance: 74% *Remission:* Change in PF	*Response* (ΔPF): AUC: 0.85 Sensitivity: 70% Specificity: 85% PPV: 0.95 NPV: 0.76 *Remission* (ΔPF): AUC: 0.66 Sensitivity: 65%

			remitters ≤10 MADRS at posttreatment		Cordance: 70% Change in MRF Cordance: 51%	Specificity: 74% PPV: 65% NPV: 74% *Response* (ΔMRF): AUC: 0.80 Sensitivity: 70% Specificity: 95% PPV: 95% NPV: 76% *Remission* (ΔMRF): AUC: 0.59 Sensitivity: 93% Specificity: 31% PPV: 39% NPV: 91%
Jaworska, 2019	51 patients with MDD	12-week double-blinded trial of: 1) escitalopram + bupropion 2) escitalopram + placebo 3) bupropion + placebo	Responders vs. Non-responders Responders defined as ≥50% reduction in MADRS scores from baseline to posttreatment	RF SVM AdaBoost CART MLP GNB	88% *Combined model, accuracy of each individual model not reported*	AUC: 0.716-0.901 *Highest AUC observed in Random Forest Model* *Combined model* Sensitivity = 77% Specificity = 99% PPV = 99 NPV = 81
Mumtaz, 2017	34 patients with MDD	Open-label trial of an SSRI	Responders vs. Nonresponders Responders defined as ≥50% improvement in pre- vs. post-treatment	LR	87.5%	Sensitivity: 95% Specificity: 80%

			BDI-II scores			
Rajpurkar, 2020	518 patients with MDD	Patients randomised in a 1:1:1: ratio to escitalopram, sertraline, or extended-release venlafaxine for 8 weeks	Regression model Predict improvement in individual symptoms, defined as the difference in score for each of the symptoms on the HAM-D from baseline to week 8.	GBM	R2 0.375-0.551 *Best model observed using EEG and baseline symptom features*	95% CI: 0.473-0.639 Used C-index to assess performance (probability that the algorithm will correctly identify, given 2 random patients with different improvement levels, which patient showed greater improvement
Wu, 2020	309 patients with MDD	8-week course of sertraline or placebo	Regression model Used pre- minus post-treatment difference in HAMD17 scores, with missing endpoint values imputed to maintain an intent-to-treat framework.	SELSER *Algorithm developed in the current study*	R^2 0.60 *Sertraline* R^2 0.41 *Placebo*	NA
Zhdanov 2020	122 patients with MDD	8-weeks of open-label escitalopram (10-20 mg) treatment	Responders vs Non-responders Responders defined as ≥50% improvement in MADRS scores from baseline to post-treatment	SVM *Radial kernel*	79.2% *Using baseline EEG data* *82.4%* *Using baseline and week 2 EEG data*	*Baseline Model* Sensitivity - 67.3% Specificity - 91.0% *Baseline and Week 2 Model* Sensitivity: 79.2% Specificity: 85.5%

Table 1: Machine learning studies predicting treatment response using EEG in major depressive disorder
Abbreviations:
ANN, Artificial Neural Network; *BDI*, Beck Depression Inventory; *BPNN*, Back-Propagation Neural Networks; *CART*, Classification and Regression Trees; *CNN*, Convolutional Neural Network; DLPFC, Dorsolateral Prefrontal Cortex; DRBMC, Discriminative Restricted Boltzmann Machine; *DT*; Decision Trees; *ELM*, Extreme Learning Machine; *GBM*, Gradient Boosting Machine, *GNB*, Gaussian Naive Bayes; *HAM-D*, Hamilton Depression Rating Scale; *IDS-SR*, Inventory of Depressive Symptomatology (Self-Report); *kNN*, k-Nearest Neighbours; *KPLSR*, *Kernelized Partial Least Squares Regression; LASSO*, least absolute shrinkage and selection operator; *LDA*, Linear Discriminant Analysis; *LR*, Logistic Regression; *MADRS*, Montgomery-Asberg Depression Rating Scale; *MFA*, Mixture of Factor Analysis; *MLP*, Multi-Layer Perceptron*; MRF*, Middle Right Frontal*; NMSC*, nearest mean classifier; *PARZEN*, Parzen density estimation; *PERCL*, perceptron classifier; *RF*, Random Forest; *SCZ*, Schizophrenia; *SELSER*, Sparse EEG Latent SpacE Regression; *SVM*, Support Vector Machine

First author, year	Pre-processing Strategy	EEG Features	Feature Extraction Method	Feature Selection Method	Top Features *Top 10 features, if applicable*
STUDIES PREDICTING RESPONSE TO NEUROSTIMULATION THERAPY					
Bailey, 2017	Data downsampled to 1000 Hz Second order Butterworth filtering with bandpass from 1-80 Hz and a band-stop filter 47-53 Hz	Power Spectral Analysis Connectivity Analysis	<u>Power Spectral Analysis</u> • *Morlet Wavelet* transform to calculate power in the upper alpha band (10-12.5 Hz), theta band (4-8 Hz), and gamma band (30-45 Hz) • Average power calculated across entire retention period with each frequency band and averaged over	*Not applicable*	*Statistically significant variables between responders and non-responders; authors did not report top features in the total model* - Greater theta power at Fz in responders vs non-responders (F1 = 8.577, p = 0.006) - No significant differences for alpha

	Fast ICA used to manually select and remove eye blinks, movements, and remaining muscle artifacts.		trials <u>Connectivity Analysis</u> • *Hanning taper time-frequency* transform to determine instantaneous phase values for complex Fourier-spectra from 4-45 Hz with a 1 Hz resolution across a 3-oscillation sliding time window • *Weighted phase lagged index* (wPLI) calculated between each electrode • wPLI provides a value between 0-1 for each electrode pair at each frequency and time point		or gamma power, or theta-gamma coupling - Responders showed a non-significant pattern of less gamma connectivity than non-responders at baseline (p=0.523), and greater gamma connectivity at week 1 (p=0.0836). - Responders showed significantly more theta connectivity across baseline and week 1, with both interhemispheric fronto-parietal coupling, and frontal and parietal interhemispheric coupling (overall p = 0.003).
Bailey 2018	*Same Procedure as Bailey 2017*	Power Spectral Analysis Connectivity Analysis Theta Cordance Analysis	*Power and connectivity analyses follow the same procedure as Bailey 2017* <u>Theta Cordance Analysis</u> • Absolute power values for each epoch 1-80 Hz underwent a multi-taper fast Fourier frequency transformation with a Hanning taper • Absolute power averaged across neighbouring electrode pairs • Relative power in reattributed absolute theta band calculated by dividing power in theta band by total power from 1-80 Hz • Subtracted half-maximal values from normalized absolute and relative power in theta band, and summed together for each electrode <u>iAPF Analysis</u> • Individualized alpha peak frequency averaged across F3, Fz, and F4 electrodes	*Not applicable*	*Statistically significant variables between responders and non-responders; authors did not report top features in the total model* - Greater theta connectivity in responders vs non-responders (p=0.0216, FDR p=0.030). Responders showed atypical, elevated theta connectivity, while non-responders showed typical theta connectivity, which was comparable to controls. - No main effect of theta cordance, frontal-midline theta power, or alpha power.

			• Multitaper fast Fourier frequency transformation • Gaussian distribution with least squared error fitted to electrodes in 6-14 Hz range • Peaks of distribution selected from each electrode and averaged		
Corlier, 2019	ICA-based FASTER algorithm Dominant alpha frequency peak determined for each subject (highest spectral peak within 7-13 Hz alpha range)	EEG functional connectivity measures (coherence, envelope correlation, and alpha band frequency)	<u>Functional Connectivity Measures</u> • Coherence: correlation of amplitude and phase • Envelope: correlation of amplitude • Alpha Frequency Band: similarity of the spectral waveform of the alpha band across regions	Elastic Net	Coherence & Envelope: connections in the frontal to temporo-parietal nodes Alpha frequency band: Connections between the left frontal seeds (near stimulation site) and contralateral fronto-temporal locations EN models for coherence and envelope correlation showed a diffuse coupling pattern, while αSC showed a more focal connectivity.
Erguzel 2014	Manually selected artifact-free EEG data with a minimum split-half reliability ratio of 0.95 and minimum test-retest reliability ratio of 0.90. FFT	EEG Cordance (combines absolute and relative EEG power, and negative discordance values)	<u>EEG Cordance</u> • Normalized power across electrode sites and frequency bands • Maximum absolute and relative power of each frequency band is calculated to derive normalized absolute and relative power • Half-maximal value is subtracted, absolute/relative normalized power is summed.	<u>Genetic Algorithm</u> • adaptive heuristic search algorithm was applied to features of all selected channels to reduce the number of dimensions	Fp1, Fp2, F7, F8, and F3 in the theta frequency band
Erguzel 2015	Band-pass filter with 0.15-30 Hz frequency FFT used to calculate absolute and relative power in each of two non-overlapping frequency bands (delta – 1-4 Hz, theta – 4-8 Hz)	EEG Cordance (combines absolute and relative EEG power, and negative discordance values)	<u>EEG Cordance</u> • Normalized power across electrode sites and frequency bands • Maximum absolute and relative power of each frequency band is calculated to derive normalized absolute and relative power • Half-maximal value is subtracted,	ANN	NA

			absolute/relative normalized power is summed.		
Erguzel 2016	Band-pass filter with 0.15-30 Hz frequency Manually selected artifact-free EEG data (at least 2 min) FFT	EEG Cordance (combines absolute and relative EEG power, and negative discordance values)	*EEG Cordance analyses follow the same procedure as Erguzel 2014*	*Not applicable*	*Feature set was composed of frequency bands for six frontal electrodes (Fp1, Fp2, F3, F4, F7 and F8)*
Hasanzadeh 2019	Sampling frequency 500 Hz Bandpass FIR filter (1-42 Hz) ICA to remove noisy data MARA to label noisy ICs Visually inspected to eliminate remaining artifacts	21 features in four categories (nonlinear, PSDl, bispectral, and cordance)	*Nonlinear Features* • LZC: Complexity measure of time series to estimate scholastic and chaotic behavior of time series • KFD: Algorithm for computing fractal dimension, a measure of self-similarity of a time series based on number of pattern repetitions *Power Spectral Density* • Delta (1-4 Hz) - Beta (12-30 Hz) by Welch method with a non-overlapped window, 500 samples in length • Average power computed for frequencies in each band *Bispectrum features* • Method that quantifies the degree of phase coupling between components of a signal *Cordance* • measure of complexity of system based on chaos and time delay reconstruction theory	*mRMR*	- Nonlinear (LZC, KFD, CD) - 80.4% accuracy - Power (D, T, A, B) - 91.3% accuracy - Bispectrum (BispSL, Bisp2M, and BispEn in all bands) - 84.8% accuracy - Cordance (Fr, Pre, Fr) - 76.1% accuracy - All - 87% accuracy
STUDIES PREDICTING RESPONSE TO PHARMACOLOGICAL TREATMENT					
Cao, 2019	Real-time artifact removal algorithm based on CCA, feature extraction, and a GMM used to improve signal	Power Spectral Analysis	<u>Power Spectral Analysis</u> • 256-point FFT using Welch's method	p-value: measured using the Wilcoxon rank-sum test with a significant p-value < 0.05.	0.5 mg/kg dose - AF7 theta - p= 0.042 - Fp2 theta - p= 0.035

	quality	EEG Alpha Asymmetry EEG Theta Cordance	• 10 min spans of data with 256-point moving window at 128-point overlap • Absolute and relative power of four prefrontal channels from delta (1-3.5 Hz), theta (4-7.5 Hz), lower alpha (8-10 Hz) and upper alpha (10.5-12 Hz) bands. <u>EEG Alpha Asymmetry</u> • mid-prefrontal (Fp1/Fp2) and mid-lateral (AF7/AF8) hemispheric asymmetry index to establish a relative measure of the difference in EEG (lower and upper) alpha power between the right and left forehead areas. <u>EEG Theta Cordance</u> • Combines information from both absolute and relative powers in the EEG theta band		0.2mg/kg dose - Fp1 theta - p= 0.038 - Fp2 theta - p= 0.042
Cooks 2020	Artifact-free epochs selected following rejection of muscle, electrocardiographic, and drowsiness artifacts.	Power Spectral Analysis ATR Relative combined theta and alpha power	<u>Power Spectral Analysis</u> • Calculated using consecutive two-second epochs of eyes-closed rest, by averaging values calculated separately for each channel in each epoch <u>Relative combined theta and alpha power</u> • Non-linear weighted combination of relative combined theta and alpha power (3-112 Hz), alpha1 power (8.5-12 Hz) and alpha2 absolute power (9-11.5 Hz)	Relative combined theta and alpha power was scaled to a range from 0-100; a cut-off score of ≥ 46.2 was selected	NA
Jaworska 2019	Bandpass filters 0.1-80 Hz 100s of artifact-free data subjected to a FFT ln-transformed prior to	eLORETA analysis Theta Cordance	*eLORETA analysis* • estimates neural activity as current density based on MNI-152 template, creating a low-resolution activation image *Theta Cordance*	*Tree-Based Feature Selection* *kernel PCA*	eLORETA features were most important, comprising 17 delta, 20 theta, 14 alpha$_1$, 20 alpha$_2$, and 17 beta EEG features. *Delta* Power at week 1 at T8 followed by

	analyses to ensure normality *(minimizes influence of extreme values)*		• Values from prefrontal electrodes (Fp1, Fp2) at baseline and week 1		power at Cp6 *Theta* Baseline power at Fp2 and week 1 power at Fc2 Alpha[1] Baseline power at F7/8 Alpha[2] Baseline power at P8 and week 1 power at O1 *Beta* Baseline power at T7 and week 21 power at Fz
Mumtaz, 2017	Bandpass filters 0.1-70 Hz EEG data collected during 5 min eyes open and 5 min eyes closed - 3-stimulus visual Oddball task used 50 Hz notch filter used to suppress power line noise	wavelet coefficients in the delta and theta frequency range	*Wavelet coefficients* • involves a window function to capture both low and high frequency components of the signal	rank-based feature selection according to their relevance to class labels minimum redundancy and maximum relevance	*Top EEG Features:* Fp2 - delta frequency C3 - theta frequency F7 - delta frequency F3 - delta frequency F7 - theta frequency T4 - theta frequency F8 - theta frequency F4 - delta frequency Fz - delta frequency F4 - delta frequency C4 - delta frequency F8 - theta frequency T4 - delta frequency P3 - theta frequency
Rajpurkar 2020	Raw EEG signal was filtered using a band-pass filter with 0.15 - 30 Hz frequency prior to artifact removal FFT	Relative and Absolute Band Power Frontal alpha asymmetry Occipital asymmetry	*Relative/Absolute Power as described above* *Frontal alpha asymmetry* • difference in alpha bandpower between O2 and O1 *Occipital beta asymmetry* • difference in beta bandpower between O2 and O1	Gradient Boosted Feature Selection	*Top EEG Features:* 1. T7-T3 alpha absolute ratio 2. T7-T3 beta absolute ratio 3. F7 gamma relative 4. Fp2 delta relative 5. F3 alpha absolute 6. Fp2 theta absolute 7. P4 alpha absolute 8. T7-T3 beta relative ratio

		Ratio of Beta/Alpha band power Ratio of Theta/Alpha band power	Ratio of Beta/Alpha and Theta/Alpha band power • Calculated for each electrode Feature Selection: *Decision Tree weight in LightGBM*		9. F7 beta relative
Salle 2020	Data was filtered (0.1-30 Hz), ocular-corrected, and inspected for artifacts (voltages ± μV, faulty channels, drift) Minimum of 100 seconds of artifact-free data was required for participant inclusion	Theta Cordance (Prefrontal – Fp1, Fp2 MRF – Fz, Fp2, F4, F8)	EEG Theta Cordance Combines information from both absolute and relative powers in the EEG theta band	NA	*Top EEG Features:* Change in prefrontal theta cordance (Fp1+ Fp2) = 81% accuracy Change in MRF theta cordance (Fz, Fp2, F4, F8) = 74% accuracy
Wu 2020	60 Hz AC line noise artifact removed using CleanLine - Non-physiological slow drifts in EEG recordings were removed using 0.01 Hz high-pass filter - Spectrally filtered EEG data were re-referenced to common average - Bad channels were rejected based on thresholding spatial correlations among channels - Subjects with more than 20% bad channels were discarded - Rejected channels were interpolated from EEG of adjacent channels via spherical spline interpolation	SELSER Channel-level alpha band power Theta Coherence Band power features of latent signals extracted with ICA or PCA	*Alpha band power and theta coherence as described above* *SELSER* • spatial filter transforms multi-channel EEG data into a single latent signal, where the power is used as a feature • model fitting is done under a sparse constraint on the number of spatial filters, which reduces dimensionality *Latent signals extracted with ICA or PCA* • eigenvalues of the covariance matrix to reduce dimensionality	SELSER	Best performance using SELSER on alpha frequency range eyes-open rsEEG data *(feature importance was not reported)*

	- Remaining artifacts were removed using ICA - EEG data re-referenced to common average				
Zhdanov 2020	0.05 - 100 Hz bandpass filter Filtering performed using 2nd order Butterworth filters applied to the data in forward and reverse direction, to eliminate phase distortion Data pre-processed with EEGLAB toolbox Channels contaminated by large sporadic artifact were identified by human analyst and deleted EEG data bandpass filtered 1-80 Hz notch-filtered at 60 Hz	Electrode-level spectral features Source-level spectral features Multiscale-entropy-based features Microstate-based features	*Electrode-level spectral features* • EEGLAB function *spectopo* to obtain power spectrum • log-transformed absolute power obtained for each channel • For each pair, absolute power at left electrode divided by right, resulting in 25 features for each band *Source-level spectral features* • eLORETA algorithm as implemented by LORETA-KEY software • Following regions selected on basis of prior literature: ACC, rACC, and mOFC *Multiscale-entropy-based features* • Quantifies variability of time series by estimating predictability of amplitude patterns across a time series • Two consecutive data points were used for data matching, and points were considered to match if their absolute amplitude difference was <15% of the standard deviation of the time series. *Microstate-based features* • Implemented using CARTOOL • average duration: average amount of time a microstate class remains stable when it appears (in	Unpaired 2-tailed t test	MSE asymmetry features - C3/C4 (baseline) MSE asymmetry features - FC3/FC4 (baseline) MSE asymmetry features - T7/T8 (week 2) MSE asymmetry features - CP3/CP4 (week 2) Electrode-level spectral asymmetry - P3/P4 alpha low (baseline) Electrode-level spectral asymmetry - T7/TP8 theta (week 2) Electrode-level spectral asymmetry - F7/F8 beta mid (week 2) Source-level spectral features - alpha high ACC, rACC (week 2)

			milliseconds) • frequency: occurrence of each microstate class per second • coverage: % of recording covered by each microstate class		

Table 2: Extracted Features Across Studies

ACC, Anterior Cingulate Cortex; *rACC*, rostral Anterior Cingulate Cortex; *ANN*, Artificial Neural Network; *CCA*, Canonical Correlation Analysis; *Coh*, Coherence; *eLORETA*, Exact low resolution brain electromagnetic tomography; *FDR*, Fisher's Discriminant Ratio; *FIR*, Finite Impulse Response; *FFT*, Fast Fourier Transformation; *GMM, Gaussian Mixture Model; ICA*, Independent Component Analysis; *KFD*, Katz Fractal Dimension; *LASSO*, Least Absolute Shrinkage and Selection Operator; *LCMV*, linearly constrained minimum variance; *LightGBM*, Light Gradient Boosting Machine; *LZC*, Lempel-Ziv Complexity; *MARA*, Multiple Artifact Rejection Algorithm; *MNI*, Montreal Neurological Institute; *mOFC*, medial Orbitofrontal Cortex; *MRF*, Middle Right Frontal; *mRMR*; Maximum Relevance Minimum Redundancy; *MSC*, Magnitude Squared Coherence; *PCA*, Principal Component Analysis; *PSD*, Power Spectral Density; *rACC*, rostral Anterior Cingulate Cortex; *rsEEG*, Resting-state EEG; *SELSER*, Sparse EEG Latent SpacE Regression

a)

Authors	Sensitivity	2.5%	97.5%	Specificity	2.5%	97.5%
Bailey, 2017	0.731	0.460	0.896	0.946	0.798	0.988
Bailey, 2018	0.700	0.448	0.870	0.914	0.758	0.973
Corlier 2019	0.607	0.494	0.709	0.643	0.477	0.780
Erguzel, 2015	0.919	0.772	0.975	0.827	0.643	0.927
Erguzel, 2016	0.841	0.665	0.945	0.938	0.769	0.985
Hasanzadeh, 2019	0.854	0.665	0.945	0.938	0.769	0.985
Cao, 2019	0.794	0.558	0.922	0.886	0.694	0.964
Cook, 2020	0.731	0.576	0.845	0.542	0.383	0.692
Salle, 2020	0.696	0.511	0.834	0.929	0.741	0.983
Jaworska, 2019	0.768	0.585	0.886	0.980	0.834	0.998
Mumtaz, 2017	0.921	0.719	0.982	0.763	0.539	0.899
Zhdanov, 2020	0.791	0.666	0.878	0.846	0.742	0.913
AVERAGE	0.776	0.600	0.892	0.846	0.678	0.923
Test for equality of sensitivities: **X-squared** = 23.09, ***p*-value** = 0.017						
Test for equality of specificities: **X-squared** = 46.23, ***p*-value** = 0.00000294						
Correlation of sensitivities and false positive rates: **Rho** = -0.203 (-0696-0.420)						
Total DOR: 23.49 (95% CI: 10.40-52.02), τ2=1.395 (95% CI: 0.00-2.13)						
posLR: 5.232 (95% CI: 3.15-8.67), τ2= 0.502 (0.00-1.24)						

negLR: 0.271 (95% CI: 0.195-0.376), τ2= 0.190 (0.00-0.495)
AUC: 0.850 (95% CI: 0.747-0.890); **pAUC**: 0.777

b)

Authors	**Mean Accuracy**	**95% CI**	**%W (random)**
Bailey, 2017	91.0%	81.34-100%	8.3%
Bailey, 2018	86.6%	82.23-91.16%	12.5%
Corlier 2019	68.5%	59.96-78.24%	7.1%
Erguzel, 2015	89.0%	80.49-98.59%	9.0%
Erguzel, 2016	86.4%	80.86-92.31%	11.5%
Hasanzadeh, 2019	91.3%	82.57-100%	9.1%
Cao, 2019	81.3%	70.04-94.36%	6.3%
Cook, 2020	64.4%	53.89-76.94%	5.0%
Salle, 2020	80.9%	69.79-93.91%	6.3%
Jaworska, 2019	88.2%	79.05-98.49%	8.4%
Mumtaz, 2017	87.50	75.77-100%	6.5%
Zhdanov, 2020	82.4%	75.58-89.83%	10.0%
Random effects model Mean = 83.93% (95% CI: 78.90-89.29)			

Table 3 - Model Performance Metrics Across EEG Models

A summary of performance metrics across all predictive models of treatment response using EEG.

A) The madad function in the "mada" package was used to calculate the sensitivity, specificity, and partial Area-Under-The-Curve (AUC) across studies, while the maduani function was used to calculate the Diagnostic Odds Ratio (DOR), positive likelihood ratio (posLR), and negative likelihood ratio (negLR). AUC was calculated using the AUC_boot function in dmetatools, with an alpha of 0.95 and 2000 bootstrap iterations. Overall, the balanced accuracy (sensitivity + specificity / 2) was 81.1%.

B) The metamean function in the "meta" package was used to pool accuracy across studies in a random effects model using an inverse variance method with Knapp-Hartung adjustments to calculate the confidence interval around the pooled effect. Across models, overall model accuracy was 83.93% (95% CI: 78.90-89.29).

Supplementary Table S1 – Machine learning studies predicting treatment response using EEG in Major Depressive Disorder (excluded studies)

First author, year	Sample size and diagnosis [1,2]	Intervention	Outcome	Machine learning model	Accuracy	Other measures
STUDIES PREDICTING RESPONSE TO NEUROSTIMULATION THERAPY						
Al-Kyasi, 2016	10 patients with MDD	15 sessions of tDCS over 3 weeks	Responders vs. Nonresponders Responders defined as ≥ 50% decrease in MADRS scores after session 15 or 23 of tDCS	SVM ELM LDA	76% *Performance was averaged across all algorithms*	N/A
Zandvakili, 2019	29 patients with comorbid MDD and PTSD	33 sessions of 5 Hz left DLPFC rTMS	Responders vs. Nonresponders Responders defined as ≥50% decrease in IDS-SR scores from baseline to end of treatment	LASSO SVM	LASSO 73-80.5% SVM 74-78.6%	MDD AUC: 0.83 Sensitivity: 47-94% Specificity: 0-83% PTSD: AUC: 0.71 Sensitivity: 37-100% Specificity: 0-100% *Sensitivity and specificity of SVM model not reported*
STUDIES PREDICTING RESPONSE TO PHARMACOLOGICAL TREATMENT						
Khodayari-Rostamabad, 2013	22 patients with MDD	Open-label trial of SSRI antidepressant	Responders vs. Nonresponders Responders defined as ≥30% improvement between the pre- and post-treatment HAMD-17 scores.	MFA	87.9%	Sensitivity: 94.9% Specificity: 80.9%
Rabinoff, 2011	25 patients with MDD	8-week double-blinded trial of either: 1) fluoxetine, 2) venlafaxine or 3) placebo	Responder vs Nonresponder Responders defined as post-treatment HAM-D scores ≤10 points	CART	*Venlafaxine* Balanced Accuracy: 91.5% *Fluoxetine* Balanced Accuracy: 85.5%	<u>*Venlafaxine*</u> Sensitivity: 83% Specificity: 100% PPV: 100% NPV: 86% <u>*Fluoxetine*</u> Sensitivity: 71% Specificity: 100% PPV: 100%

						NPV: 75%
Shahabi, 2021	30 patients with MDD	4-week course of an SSRI	Responders vs Nonresponders Responders defined as ≥50% improvement in BDI-II scores from baseline to post-treatment	CNN	95.74%	Sensitivity: 95.56% Specificity: 95.64%

BDI, Beck Depression Inventory; *CNN*, Convolutional Neural Network; DLPFC, Dorsolateral Prefrontal Cortex; *ELM*, Extreme Learning Machine; *GBM*, Gradient Boosting Machine, *HAM-D*, Hamilton Depression Rating Scale; *IDS-SR*, Inventory of Depressive Symptomatology (Self-Report); *kNN*, k-Nearest Neighbors; *LASSO*, least absolute shrinkage and selection operator; *LDA*, Linear Discriminant Analysis; *LR*, Logistic Regression; *MADRS*, Montgomery-Asberg Depression Rating Scale; *MFA*, Mixture of Factor Analysis; *PARZEN*, Parzen density estimation; *RF*, Random Forest; *SVM*, Support Vector Machine

First author, year	EEG System	Reference Choice	Impedance	Filtering Method	Electrooculogram used?	Electrocardiogram used?	Eyes Open (EO) Eyes Closed (EC)
STUDIES PREDICTING RESPONSE TO NEUROSTIMULATION THERAPY							
Bailey, 2017	30-channel Ag/AgCl electrode EasyCap EEG system	CPz	<5 kΩ	Bandpass filter (1-80 Hz) Bandstop filter (47-53 Hz)	No	No	EC
Bailey, 2018	30-channel Ag/AgCl electrode EasyCap EEG system	CPz	<5 kΩ	Bandpass filter (1-80 Hz) Bandstop filter (47-53 Hz)	No	No	EO/EC
Corlier, 2019	64-channel ANT Neuro TMS-compatible EEG system	CPz	<10 kΩ	Bandpass filter (0.5-55 Hz)	Yes	No	NA
Erguzel, 2014	19-channel Scan LT EEG amplifier and electrode cap (6 channels were used)	Linked Ears M1 + M2, LE, RE	NA	Bandpass filter (0.15-30 Hz)	No	No	EC
Erguzel, 2015	19-channel Scan LT EEG amplifier and electrode cap	Linked Ears M1 + M2, LE, RE	NA	Bandpass filter (0.15-30 Hz)	No	No	EC
Erguzel, 2016	19-channel Scan LT EEG amplifier and electrode cap	Linked Ears M1 + M2, LE, RE	NA	Bandpass filter (0.15-30 Hz)	No	No	EC
Hasanzadeh, 2019	Mitsar-EEG 201 18 Ag/AgCL electrodes	Linked Ears M1 + M2, LE, RE	NA	Bandpass filter (1-42 Hz)	No	No	EC
STUDIES PREDICTING RESPONSE TO PHARMACOLOGICAL TREATMENT							
Cao, 2019	Mindo-4S Jellyfish	A2	NA	Bandpass filter (1-12 Hz)	No	No	EC

	4 dry electrodes (Fp1, Fp2, AF7, AF8)						
Cook, 2020	Covidien BIS Complete 4-Channel Monitor 4 channel system (FPz, FT7, FT8, A1/A2)	A1+A2	NA	NA	No	No	NA
De la Salle, 2020	32 channel EasyCap EEG with Ag/AgCl electrodes	Common Average Reference	≤5 kΩ	Bandpass filter (0.1-30 Hz)	Yes	No	EC
Jaworska, 2019	32 channel EasyCap EEG with Ag/AgCl electrodes	Common Average Reference	≤5 kΩ	Bandpass filter (0.1-30 Hz)	Yes	No	EC
Mumtaz, 2017	19 channel electro-gel sensors with linked ear references - Brain Master Discovery amplifier was used	Linked Ear Reference	NA	Bandpass filter (0.1-70 Hz)	No	No	EC/EO
Rajpurkar, 2020	Scan LT EEG amplifier and electrode cap 6 frontal electrodes used *(Fp1, Fp2, F3, F4, F7, an F8)*	NA	NA	NA	No	No	EC
Wu, 2020	*Data from four studies* BioSemi (72 channels) NeuroScan Synamp (62 channels) NeuroScan Synamp (60 channels) Geodesic Net (129 channels)	Common Average Reference	<50 kΩ	0.01 Hz high-pass filter 100 Hz low-pass filter	No	No	EC/EO
Zhdanov, 2020	*Data from four sites* 58 electrodes	Common Average Reference	NA	Bandpass filter (1 - 80 Hz) Notch-filtered at 60 Hz	No	No	EC

Supplementary Table S2 – Characteristics of EEG Systems

Study	RISK OF BIAS	APPLICABILITY CONCERNS

	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Bailey, 2017	☺	☺	☹	☺	☺	☺	☺
Bailey, 2018	?	☺	☹	☺	☺	☺	☺
Cao, 2019	☹	☺	☺	☺	☺	☺	☺
Cook, 2020	☺	?	?	☺	☺	☺	☺
Corlier, 2019	☺	☺	☹	☺	☺	☺	☺
De la Salle, 2020	☺	?	☹	☺	☺	☺	☺
Erguzel, 2014	☺	☺	☹	☺	☺	☺	☺
Erguzel, 2015	☹	?	☹	☺	☺	☺	☺
Erguzel, 2016	☺	☺	☹	☺	☺	☺	☺
Hasanzadeh, 2019	☺	?	☹	☺	☺	☺	☺
Jaworska, 2019	☺	☺	☺	☺	☺	☺	☺
Mumtaz, 2017	☺	☺	☹	☺	?	☺	☺
Rajpurkar, 2020	☺	☺	☺	☺	☺	☺	☺
Wu, 2020	☺	☺	☺	☺	☺	☺	☺
Zhdanov, 2020	☺	☺	☺	☺	☺	☺	☺

☺ Low Risk ☹ High Risk ? Unclear Risk

Supplementary Table S3 – Quality Assessment of Diagnostic Accuracy Studies-2 (QUADRS-2)

Authors	Classification Task	Method to address class imbalance	True and False Positive/Negative	Performance Metrics	95% Confidence Intervals of Accuracy

Bailey, 2017	Responders (≥ 50% improvement in HAMD-17) vs *Non-responders* (10/29)	N/A	TP = 9 FP = 1 TN = 26 FN = 3	Balanced Accuracy = 91% Sensitivity = 90% Specificity = 92% False Positive = 8% False Negative = 10% Standard Error = 5.20 Standard Deviation = 32.47	Accuracy = 91% (95% CI: 77.36-97.76)
Bailey, 2018	*Responders* (> 50% improvement in HAMD-17) vs *Non-responders* (12/30)	Class weights	TP = 10 FP = 2 TN = 26 FN = 4	Balanced Accuracy = 86.50% Sensitivity = 84% Specificity = 89% False Positive = 11% False Negative = 16% Standard Error = 2.27 Standard Deviation = 12.84	Accuracy = 86.60% (95% CI: 82.14-91.06)
Corlier, 2019	*Responders* (≥ 49% improvement in IDS-30) vs *Non-responders* (68/41)	N/A	TP = 45 FP = 12 TN = 22 FN = 29	Balanced Accuracy = 69% Sensitivity = 67.1% (19.2) Specificity = 70.9% (13.3)	Accuracy = 68.50% (95% CI: 58.86-77.10)

				False Positive = 29.1% False Negative = 32.9% Standard Error = 4.65 Standard Deviation = 48.54	
Erguzel, 2014	*Responders* (≥50% improvement in HAMD-17) vs *Non-responders* (90/57)	N/A	Not available	Balanced Accuracy = N/A Sensitivity = 84.44% Specificity = N/A False Positive = N/A False Negative = 15.56% Standard Error = N/A Standard Deviation = N/A	Accuracy = 80.25%
Erguzel, 2015	*Responders* (≥50% improvement in HAMD-17) vs *Non-responders* (30/25)	N/A	TP = 28 FP = 4 TN = 21 FN = 2	Balanced Accuracy = 88.66% Sensitivity = 93.33% Specificity = 84.00% False Positive = 16% False Negative =	Accuracy = 89.09% (95% CI: 77.85-95.94)

				6.7% Standard Error = 4.61 Standard Deviation = 34.18	
Erguzel, 2016	*Responders* (≥50% improvement in HAMD-17) vs *Non-responders* (90/57)	N/A	TP = 76 FP = 5 TN = 52 FN = 14	Balanced Accuracy = 87.70% Sensitivity = 84.30% Specificity = 91.11% False Positive = 8.8% False Negative = 15.7% Standard Error = 2.92 Standard Deviation = 35.40	Accuracy = 86.4% (95% CI: 80.56-92.04)
Hasanzadeh, 2019	*Responders* (≥50% improvement in HAMD-24) vs *Non-responders* (23/23)	N/A	TP = 20 FP = 1 TN = 22 FN = 3	Balanced Accuracy = 91.3% Sensitivity = 87% Specificity = 95.7% False Positive = 4.3% False Negative = 13% Standard Error = 4.68	Accuracy = 91.3% (95% CI: 79.21-97.58)

				Standard Deviation = 31.74	
Cao, 2019	*Responders* (≥45% improvement in HAMD-17) vs *Non-responders* (16/21)	Oversampling minority class	TP = 13 FP = 2 TN = 19 FN = 3	Balanced Accuracy = 87% Sensitivity = 82.1% Specificity = 91.9% False Positive = 8.1% False Negative = 17.9% Standard Error = 6.18 Standard Deviation = 37.59	Accuracy = 81.3% (95% CI: 71.23-95.47)
Cook, 2020	*Remission* (≤7 HAMD-17) vs *Non-remission* (38/35)	N/A	TP = 28 FP = 16 TN = 19 FN = 10	Balanced Accuracy = 64.8% Sensitivity = 74.3% Specificity = 55.3% False Positive = 44.7% False Negative = 25.7% Standard Error = 5.85 Standard Deviation = 49.98	Accuracy = 64.4% (95% CI: 52.30-75.24)

Salle, 2020	*Responders* (≥50% improvement in MADRS) vs *Non-responders* (27/20)	N/A	TP = 19 FP = 1 TN = 19 FN = 8	Balanced Accuracy = 82.5% Sensitivity = 70% Specificity = 95% False Positive = 5% False Negative = 30% Standard Error = 6.13 Standard Deviation = 42.02	Accuracy = 80.96% (95% CI: (66.87-90.93)
Jaworska, 2019	*Responders* (≥50% improvement in MADRS) vs *Non-responders* (27/24)	N/A	TP = 21 FP = 0 TN = 24 FN = 6	Balanced Accuracy = 88% Sensitivity = 77% Specificity = 99% PPV = 99% NPV = 81% Standard Error = 4.95 Standard Deviation = 35.35	Accuracy = 88.24% (95% CI: 76.14-95.56)
Mumtaz, 2017	*Responders* ≥50% improvement in BDI-II) vs *Non-responders* (17/17)	N/A	TP = 17 FP = 4 TN = 14 FN = 1	Balanced Accuracy = 87.5% Sensitivity = 95% Specificity = 80% False Positive = 20% False Negative = 5% Standard Error =	Accuracy = 86.11% (95% CI: 70.50-95.33)

				6.33 Standard Deviation = 37.44	
Zhdanov, 2020	*Responders* (≥50% improvement in MADRS) vs *Non-responders* (55/67)	N/A	*Model 1* TP = 43 FP = 10 TN = 57 FN = 11 *Model 2* TP = 37 FP = 6 TN = 61 FN = 18	Balanced Accuracy = 79.2% Sensitivity = 67.3% Specificity = 91.0% False Positive = 9% False Negative = 32.7% Standard Error = 3.78 Standard Deviation = 41.75 Balanced Accuracy = 82.35% Sensitivity = 79.2% Specificity = 85.5% False Positive = 14.5% False Negative = 20.8% Standard Error = 3.63 Standard Deviation	Accuracy = 80.33% (95% CI: 72.12-86.97) Accuracy = 82.4% (95% CI: 74.68-88.91)

				= 40.09	

Supplementary Table S4 – Confusion Matrices of Classification Models

False positive rate is calculated as 1-specificity, while false negative is calculated as 1-sensitivity. In cases where confidence intervals were not reported, this metric was calculated using the true/false positive/negative ratios, as well as the prevalence of the positive class (responders). Standard error was imputed by subtracting the upper bound of the 95% CI from the lower bound and dividing by 3.92 (upper bound - lower bound)/3.92. Additionally, confusion matrices were provided according to the method used to address class imbalance, where applicable. It is important to note that none of the included studies reported the true positives/true negatives and false positives/false negative rates, and the numbers indicated in the table reflect calculations based on the prevalence, sensitivity, specificity, and total sample size. Summary statistics that were not reported in studies are indicated as N/A.

First author, year	Resting state EEG used?	Task-specific EEG used?	Comments
STUDIES PREDICTING RESPONSE TO NEUROSTIMULATION THERAPY			
Bailey, 2017	No	Yes	Sternberg Working Memory Task
Bailey, 2018	Yes	No	
Corlier, 2019	Yes	No	
Erguzel, 2014	Yes	No	
Erguzel, 2015	Yes	No	
Erguzel, 2016	Yes	No	
Hasanzadeh, 2019	Yes	No	
STUDIES PREDICTING RESPONSE TO PHARMACOLOGICAL TREATMENT			
Cao, 2019	Yes	No	
Cook, 2020	Yes	No	
De la Salle, 2020	Yes	No	
Jaworska, 2019	Yes	No	
Mumtaz, 2017	No	Yes	3-stimulus visual Oddball Task
Rajpurkar, 2020	Yes	No	
Wu, 2020	Yes	No	
Zhdanov 2020	Yes	No	

Supplementary Table S5 – Resting-state and task-specific EEG

Supplementary Material

1. ML Quality Scores of All Studies

ML Quality Scores of All Studies

Predicting response to Neurostimulation										
Authors	Representative	Confounding	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Bailey, 2017	No	Yes	Yes	Yes	No	No	Yes	Yes	No	5/9
Bailey, 2018	No	Yes	Yes	Yes	No	No	Yes	Yes	No	5/9
Corlier, 2019	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	7/9
Erguzel, 2014	No	No	Yes	Yes	Yes	No	Yes	Yes	No	5/9
Erguzel, 2015	No	No	Yes	Yes	Yes	No	Yes	Yes	No	5/9
Erugzel, 2016	No	No	Yes	Yes	Yes	No	Yes	Yes	No	5/9

Hasanzadeh, 2019	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No	6/9

ML Quality Scores of All Studies

Predicting treatment response to psychiatric medication										
Authors	Representative	Confounding	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Cao, 2019	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	7/9
Cook, 2020	No	Yes	Yes	Yes	No	No	Yes	Yes	No	5/9
Jaworska, 2019	No	Yes	Yes	Yes	Yes	No	Yes	No	No	5/9
Rajpurkar, 2020	Yes	Yes	Yes	Yes	Yes	No**	Yes	Yes	No	7/9
De la Salle, 2020	No	Yes	Yes	Yes	No	No	Yes	Yes	No	5/9
Mumtaz, 2017	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No	6/9

Wu, 2020	Yes	Yes	Yes	Yes	Yes	No**	Yes	Yes	Yes	9/9
Zhdanov, 2020	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	8/9
** Class imbalance methods are not applicable to regression-based models										

2. Quality assessment instrument development

We formed a group of multidisciplinary researchers from the fields of Neuroscience, Psychiatry, and Computer Science to develop a time efficient and practical assessment strategy to evaluate the quality of supervised machine learning based healthcare research. For that purpose, we attempted to capture the reliability of the results presented in each study and identify practical ways that methodology may be improved. This instrument is not intended to provide an exhaustive evaluation of all components of supervised machine learning studies, but rather provide a brief overview of common considerations in supervised models, including patient sample, the specific outcome, algorithm selection, and how performance was evaluated. In total, this comprised nine methodological features, including sample representativeness, confounding variables, and outcome assessments Relevant considerations of each methodological feature are discussed in further detail in the next sections. The six remaining dimensions assess the quality and specific components of the machine learning approach that were used in each study. In summary, this entails the algorithm or framework used, evidence that hyper-parameter optimization and feature selection procedures were used, whether authors provided details on how missing data and class imbalance problems were handled, the accuracy of a given model, and finally whether the model performance was tested in unseen data. These dimensions were qualitatively evaluated according to the information in section 3.

3. Quality assessment instrument domains

Methodological Feature	**Considerations**
1. Representativeness of the sample	Was the study representative of the heterogeneity observed in the target population? If not, was this related to the sampling method, insufficient sample size or inclusion/exclusion criteria?
2. Confounding variables	Did the study control for the most relevant confounding variables? If so, were covariates assessed using

	subjective or objective measures?
3. Outcome assessment	How were outcome measures assessed? A. Independent blind assessment (✓) B. Secure record (e.g., surgical records) (✓) C. Interview not blinded, self-report or medical record D. No description *A-C scored as "Yes"; D scored as "No"*
4. Machine learning approach	Was the machine learning algorithm used to analyse the data clearly described and appropriate?
5. Feature selection	Did the study describe both feature selection and hyperparameter tuning? Which metrics were used?
6. Class imbalance	Did the authors address the class imbalance problem? Which method was used?
7. Missing data	Did the study describe how the authors handled missing data, including whether they were inputted or removed?
8. Performance/accuracy	Were the following performance metrics included for classification studies? 1. Accuracy 2. Sensitivity 3. Specificity 4. AUC 5. PPV/NPV Or, alternatively, were one of the following performance metrics included for regression studies? 1. Mean-squared error 2. Mean-absolute error

	3. Root-mean-squared error
9. Testing/validation	Was the test dataset "unseen" during model training? Was the model tested on a hold-out or an external dataset?

3.1. Representativeness of the sample

Machine learning models can deal with large amounts of data and the problem of heterogeneity. Therefore, there is less of a need to be restrictive with inclusion and exclusion criteria, relative to a traditional statistical approach examining significant effects at a group-level. Considering all studies included in the present review used data from randomized clinical trials, determined whether 1) performance was tested on an external sample with differences in inclusion/exclusion criteria, and 2) whether a training sample of ≥ 100 patients was used in model development.

3.2. Internal CV

To adequately control for confounding variables within machine learning models, it is important to ensure that these variables have a similar effect across the entire sample. To achieve this, randomization is an important step within the analysis. Often, the overall sample is randomly split into training and testing sets, and the analysis is repeated on the training dataset with different hyperparameters in order to maximize accuracy and minimize error. This is known as internal cross-validation. From here, if model performance is similar in the testing dataset, it presumes that potential confounding variables are uniformly distributed across the sample. Using the aforementioned criteria, we evaluated whether the authors controlled for confounding variables.

3.3. Outcome assessment

How an outcome is defined has several important implications in a predictive model. Depending on the question or problem, a classification task may be appropriate, which uses a categorical outcome, or a regression task may be more relevant, where the outcome is continuous and numeric. A clinical instrument or questionnaire, for example, can be used as a numeric score or it can be transformed into a categorical outcome by using a cut-off score. We evaluated how authors assessed these outcomes, considering (A) independent blind assessments and secure records as high quality, (B) unblinded interview, self-report, or medical record as lower quality and (C) when no description was available.

3.4. Algorithm selection

There are several algorithms to choose from, with each relying on slightly different assumptions of the underlying data. Broadly speaking, there are linear (logistic regression, linear support vector machine), non-linear (Naive Bayes, K-Nearest Neighbors, Learning Vector Quantization)

tree-based (decision trees, random forest, xgboost) and neural network (convolutional neural network, multilayer perceptrons) models, although others exist. Certain algorithms may be better suited to certain problems. For example, tree-based models such as random forest may be better suited to datasets with multicollinearity among features than linear-based models such as logistic regression. However, regularization parameters can be used in linear-based models (such as L2 regularization) to account for issues such as this.

Nevertheless, it is often difficult to determine beforehand which algorithms will lead to the highest model performance. Therefore, it is often a good strategy to compare the model performance of several algorithms. In this item, we evaluated whether the authors used an algorithm that is commonly used for the specific type of dataset, if several algorithms were compared, and if hyperparameter tuning was used.

The appropriateness of a machine learning algorithm was determined based on whether the specific data used in model development was congruent or incongruent with the strengths and limitations of the specific algorithm. For example, if a Gaussian process model was used, which is a non-sparse algorithm that loses efficiency in high dimensional spaces, in conjunction with a high-dimensional dataset, this algorithm would be deemed inappropriate for the input data. Conversely, Naive Bayes, which works well with high dimensional data would be considered an appropriate algorithm in such cases. Another example of an inappropriate model would be the use of convolutional neural networks for structural and tabular style datasets, as such algorithms are better suited to unstructured datasets. In cases where authors included both appropriate and inappropriate algorithms during model development, this consideration is scored with a “B”, alongside an asterisk to indicate which algorithms were inappropriate and why. Studies which only utilized one algorithm during model development that was deemed inappropriate received a score of “C”. Furthermore, studies are scored with a “B” if they did not compare multiple algorithms during model development and were scored as an “A” if they compared multiple algorithms that were deemed appropriate based on the candidate feature set.

3.5. Feature selection

A common problem in machine learning studies is the so-called small-n-large-p problem, also known as the curse of dimensionality, which occurs when there are more variables than examples in a dataset. Machine learning models created using these datasets are more prone to overfitting, which often results in overinflated performance in a training dataset, but much poorer performance in an external testing dataset. In addition, some algorithms cannot deal with more dimensions than examples. Highly correlated variables can also introduce more importance to a specific characteristic, decreasing the importance of the remaining variables. To circumvent these issues, a proper feature selection procedure, when applicable, should be done prior to training or as part of the training procedure, such as it happens in embedded methods. The feature selection can be knowledge-driven or data-driven. In this item, we examined if the study used a proper feature selection (if applicable).

3.6. Class imbalance

Class imbalance occurs when the distribution of the outcome classes is highly unbalanced, i.e., when one outcome occurs much more frequently than the other outcome(s). This may result in a model with high accuracy but with very little clinical utility. For example, let us suppose that we have 95 occurrences of response in our dataset and only 5 occurrences of a nonresponse. Even if our model has 95% accuracy, it is useless if the model cannot detect the five instances of non-response high accuracy. In this item, we evaluated whether there was a class imbalance in the sample and if this problem was correctly addressed. This can be done using a series of methods, including (1) changing the metric of performance (accuracy, for example, is a poor form of evaluating imbalanced data sets; (2) resampling the data set by artificially increasing it (oversampling) or by removing examples from the majority class to create a more balanced data set (undersampling); (3) by generating more data with algorithms such as the Synthetic Minority Over-Sampling Technique (SMOTE); (4) by choosing algorithms that deal better with unbalanced classes, such as CART or random forests; (5) by using penalized models; or (6) by using anomaly and change detection. In cases where class imbalance was not relevant (balanced classes or regression models) this is scored as "yes".

3.7. Missing data

It is critical to handle missing data since several algorithms cannot process incomplete data sets. Furthermore, it is also necessary to use an adequate imputation method to avoid introducing bias, which would otherwise lead to false conclusions if not addressed. It is important to report the amount of missing data in each variable, if these cases were excluded, or if the authors used an algorithm to input data and which algorithm/technique was used. Ideally, authors should provide a visual distribution of the patterns of missing data, such as aggregation plots, spinogram/spineplots, mosaic plots, etc. All these factors were evaluated in this section.

3.8. Performance/accuracy

Here, we evaluate whether the authors reported all relevant results and if they used the appropriate metrics. Studies informing only partial metrics may mask bias and flaws of the method, preventing the reader from fully understanding the relevance of the model. Confidence intervals should ideally be available for all performance metrics.

3.9. Testing/Validation

We can divide the machine learning process into three main components: training, validation, and testing. A training set allows the algorithm to learn and develop a predictive model. The validation set contains unseen data and is used to control for overfitting. Frequently, the same dataset is divided into training and validation sets. After a model is trained and validated, and shows consistent performance in both these steps, the model can be applied in an external and independent testing set. This allows us to see if the model can be generalized outside of the original sample. Some validation methods include holdout validation, k-fold, and leave one out cross validation.

A model that shows good performance in the training set but performs significantly poorer in the validation step is most likely due to overfitting - which occurs when the model relies more on the specific nuances and noise of the training dataset, resulting in poor accuracy in unseen data. In this item, we evaluated whether the authors properly tested and validated their models by taking steps to improve its generalizability. It is important to highlight that the use of cross-validation to evaluate performance should be discouraged when the data is large enough for a training-test split. Furthermore, the size of the test set should be sufficiently large for accuracy and other metrics to be estimated with high reliability.

4. Search Filter

PubMed/MEDLINE

Abbreviated Search: ("Supervised Machine Learning" OR "Artificial intelligence") AND ("Major Depressive Disorder") AND ("Electroencephalography") AND ("Intervention" OR "Treatment")

Full Search:
((Artificial Intelligence[MeSH Major Topic]) OR (Supervised Machine Learning[MeSH Major Topic])) AND (Depressive Disorder, Major[MeSH Major Topic])) OR (Major Depressive Disorders[MeSH Terms])) OR (Major Depressive Disorder[MeSH Terms])) OR (Depressive Disorders[MeSH Terms])) OR (Neurosis, Depressive[MeSH Terms])) OR (Depressive Neuroses[MeSH Terms])) OR (Depressive Neurosis[MeSH Terms])) OR (Neuroses, Depressive[MeSH Terms])) OR (Depression, Endogenous[MeSH Terms])) OR (Depressions, Endogenous[MeSH Terms])) OR (Endogenous Depression[MeSH Terms])) OR (Endogenous Depressions[MeSH Terms])) OR (Depressive Syndrome[MeSH Terms])) OR (Depressive Syndromes[MeSH Terms])) OR (Syndrome, Depressive[MeSH Terms])) OR (Syndromes, Depressive[MeSH Terms])) OR (Depression, Neurotic[MeSH Terms])) OR (Depressions, Neurotic[MeSH Terms])) OR (Neurotic Depression[MeSH Terms])) OR (Neurotic Depressions[MeSH Terms])) OR (Melancholia[MeSH Terms])) OR (Melancholias[MeSH Terms])) OR (Unipolar Depression[MeSH Terms])) OR (Depression, Unipolar[MeSH Terms])) OR (Depressions, Unipolar[MeSH Terms])) OR (Dysthmic Disorder[MeSH Terms])) OR (Disorder, Dysthymic[MeSH Terms])) OR (Dysthymic Disorders[MeSH Terms])) OR (Dysthymia[MeSH Terms])) OR (Persistent Depressive Disorder, Dysthymia[MeSH Terms])) OR (Dysthymia and Chronic Depression[MeSH Terms])) OR (Neurotic Depression, Persistent Depressive Disorder[MeSH Terms])) AND (Electroencephalography[MeSH Major Topic])) OR (EEG[MeSH Terms])) OR (Electroencephalogram[MeSH Terms])) OR (Electroencephalograms[MeSH Terms])) OR (Brain Waves[MeSH Major Topic])) AND (Clinical Trials as Topic[MeSH Major Topic])) OR (Treatment response[Other Term])) OR (treatment prediction[Other Term])) OR (treatment selection[Other Term]))

Date: 2022-02-11

Retrieved references: 1827

Scopus

Abbreviated Search: (Supervised Machine Learning OR Artificial Intelligence) AND (Major Depressive Disorder) AND (Electroencephalography) AND (Intervention OR Treatment)

Full Search: ((Artificial Intelligence) OR (Supervised machine Learning)) AND ((Depressive Disorder, Major) OR (Major Depressive Disorders) OR (Depressive Disorders) OR (Neurosis, Depressive) OR (Depressive Neurosis) OR (Neuroses, Depressive) OR (Depression, Endogenous) OR (Depressions, Endogenous) OR (Endogenous Depression) OR (Endogenous Depressions) OR (Depressive Syndrome) OR (Depressive Syndromes) OR (Syndrome, Depressive) OR (Syndromes, Depressive) OR (Depression, Neurotic) OR (Depressions, Neurotic) OR (Neurotic Depression) OR (Neurotic Depressions) OR (Melancholia) OR (Melancholias) OR (Unipolar Depression) OR (Depression, Unipolar) OR (Depressions, Unipolar) OR (Unipolar Depressions) OR (Dysthmic Disorder) OR (Disorder, Dysthymic) OR (Dysthymic Disorders) OR (Dysthymia) OR (Persistent Depressive Disorder, Dysthymia) OR (Dysthymia and Chronic Depression) OR (Neurotic Depression, Persistent Depressive Disorder)) AND ((Electroencephalography) OR (EEG) OR (Electroencephalogram) OR (Electroencephalograms) OR (Brain Waves)) AND (Clinical Trials) OR (Treatment Response) OR (Treatment Prediction) OR (Treatment Selection)

Date: 2022-02-11

Retrieved References: 1466

Web of Science

Search: (TS= Algorithms OR Machine Learning OR Artificial Intelligence) AND (TS= Major Mental Disorder) AND (TS =Electroencephalography OR Magnetoencephalography) AND (TS = Intervention OR Treatment)
Full Search:

(TS=(Artificial Intelligence) OR TS= (Machine Learning)) AND (TS=(Major Depressive Disorder) OR (TS=Depressive Disorder, Major) OR (TS=Major Depressive Disorders) OR (TS=Depressive Disorders) OR (TS=Depression) OR (TS=Dysthymia) OR (TS=Neurosis, Depressive) OR (TS=Depressive Neurosis) OR (TS=Neuroses, Depressive) OR (TS=Depression, Endogenous) OR (TS=Depressions, Endogenous) OR (TS=Endogenous Depression) OR (TS=Endogenous Depressions) OR (TS=Depressive Syndrome) OR (TS=Depressive Syndromes) OR (TS=Syndrome, Depressive) OR (TS=Syndromes, Depressive) OR (TS=Depression, Neurotic) OR (TS=Depressions, Neurotic) OR (TS=Neurotic Depression) OR (TS=Neurotic Depressions) OR (TS=Melancholia) OR (TS=melancholicas) OR (TS=Unipolar Depression) OR (TS=Depression, Unipolar) OR (TS=Depressions, Unipolar) OR (TS=Unipolar Depressions) OR (TS=Dysthmic Disorder) OR (TS=Disorder, Dysthymic) OR (TS=Dysthymic Disorders) OR (TS=Dysthymia) OR (TS=Persistent Depressive Disorder, Dysthymia) OR (TS=Dysthymia) OR (TS=Chronic Depression) OR (TS=Neurotic Depression, Persistent Depressive Disorder)) AND (TS=(Electroencephalography) OR (AB=EEG) OR (TS=Electroencephalogram) OR (TS=Electroencephalograms) OR (TS=Brain Waves)) AND (TS=(Clinical Trials) OR (TS=Treatment Response) OR (TS=Treatment Prediction) OR (TS=Treatment Selection) OR (TS=Treatment) OR (TS=Therapy))

Date: 2022-02-11

References Retrieved: 53

DUPLICATES

Retrieved References (without duplicates): 2489

Removed Duplicates: 85

References

1. de Fruyt, J. *et al.* Second generation antipsychotics in the treatment of bipolar depression: a systematic review and meta-analysis. *Journal of Psychopharmacology* **26**, 603–617 (2012).
2. Thomas, L. *et al.* Prevalence of treatment-resistant depression in primary care: cross-sectional data. *British Journal of General Practice* **63**, e852–e858 (2013).
3. Gianey, H. K. & Choudhary, R. Comprehensive Review On Supervised Machine Learning Algorithms. in *Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017* vols. 2018-January 38–43 (Institute of Electrical and Electronics Engineers Inc., 2018).
4. Osarogiagbon, A. U., Khan, F., Venkatesan, R. & Gillard, P. Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. *Process Safety and Environmental Protection* **147**, 367–384 (2021).
5. Yu, T. & Zhu, H. Hyper-Parameter Optimization: A Review of Algorithms and Applications. (2020).
6. Chekroud, A. M. *et al.* The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* **20**, 154–170 (2021).
7. Takamiya, A. *et al.* Predicting Individual Remission After Electroconvulsive Therapy Based on Structural Magnetic Resonance Imaging. *The Journal of ECT* **36**, 205–210 (2020).
8. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
9. Fingelkurts, A. A. & Fingelkurts, A. A. Altered Structure of Dynamic Electroencephalogram Oscillatory Pattern in Major Depression. *Biological Psychiatry* **77**, 1050–1060 (2015).
10. Fink, A. & Benedek, M. EEG alpha power and creative ideation. *Neuroscience & Biobehavioral Reviews* **44**, 111–123 (2014).
11. Roslan, N. S., Amin, H. U., Izhar, L. I., Saad, M. N. M. & Sivapalan, S. Role of EEG delta and beta oscillations during problem solving tasks. in *2016 6th International Conference on Intelligent and Advanced Systems (ICIAS)* 1–4 (IEEE, 2016). doi:10.1109/ICIAS.2016.7824138.
12. Amzica, F. & Steriade, M. Electrophysiological correlates of sleep delta waves. *Electroencephalography and Clinical Neurophysiology* **107**, 69–83 (1998).
13. Lally, N. *et al.* Glutamatergic correlates of gamma-band oscillatory activity during cognition: A concurrent ER-MRS and EEG study. *Neuroimage* **85**, 823–833 (2014).
14. Jacobs, G. D. & Friedman, R. EEG Spectral Analysis of Relaxation Techniques. *Applied Psychophysiology and Biofeedback* **29**, 245–254 (2004).
15. de Aguiar Neto, F. S. & Rosa, J. L. G. Depression biomarkers using non-invasive EEG: A review. *Neuroscience & Biobehavioral Reviews* **105**, 83–93 (2019).
16. Roh, S.-C., Park, E.-J., Shim, M. & Lee, S.-H. EEG beta and low gamma power correlates with inattention in patients with major depressive disorder. *Journal of Affective Disorders* **204**, 124–130 (2016).
17. Clark, D. L., Brown, E. C., Ramasubbu, R. & Kiss, Z. H. T. Intrinsic Local Beta Oscillations in the Subgenual Cingulate Relate to Depressive Symptoms in Treatment-Resistant Depression. *Biological Psychiatry* **80**, e93–e94 (2016).
18. Fitzgerald, P. J. & Watson, B. O. Gamma oscillations as a biomarker for major depression: an emerging topic. *Translational Psychiatry* **8**, 177 (2018).
19. Widge, A. S. *et al.* Electroencephalographic Biomarkers for Treatment Response Prediction in Major Depressive Illness: A Meta-Analysis. *American Journal of Psychiatry* **176**, 44–56 (2019).
20. Liberati, A. *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* **339**, b2700–b2700 (2009).
21. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of Best Practices for Evidence for Prediction. *JAMA Psychiatry* **77**, 534 (2020).
22. Whiting, P. F. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Annals of Internal Medicine* **155**, 529 (2011).

23. Doebler, P. & Holling, H. *Meta-Analysis of Diagnostic Accuracy with mada*. http://r-forge.r-project.org/projects/mada/.
24. Bailey, N. W. *et al.* Responders to rTMS for depression show increased fronto-midline theta and theta connectivity compared to non-responders. *Brain Stimulation* **11**, 190–203 (2018).
25. Bailey, N. *et al.* Differentiating responders and non-responders to rTMS treatment for depression after one week using resting EEG connectivity measures. *Journal of Affective Disorders* **242**, 68–79 (2019).
26. Corlier, J. *et al.* Changes in Functional Connectivity Predict Outcome of Repetitive Transcranial Magnetic Stimulation Treatment of Major Depressive Disorder. *Cerebral Cortex* **29**, 4958–4967 (2019).
27. Erguzel, T. T., Ozekes, S., Tan, O. & Gultekin, S. Feature Selection and Classification of Electroencephalographic Signals. *Clinical EEG and Neuroscience* **46**, 321–326 (2015).
28. Erguzel, T. T. *et al.* Neural Network Based Response Prediction of rTMS in Major Depressive Disorder Using QEEG Cordance. *Psychiatry Investigation* **12**, 61 (2015).
29. Erguzel, T. T. & Tarhan, N. Machine Learning Approaches to Predict Repetitive Transcranial Magnetic Stimulation Treatment Response in Major Depressive Disorder. in 391–401 (2018). doi:10.1007/978-3-319-56991-8_29.
30. 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. *Journal of Affective Disorders* **256**, 132–142 (2019).
31. Cao, Z. *et al.* Identifying Ketamine Responses in Treatment-Resistant Depression Using a Wearable Forehead EEG. *IEEE Transactions on Biomedical Engineering* **66**, 1668–1679 (2019).
32. Cook, I. A., Hunter, A. M., Caudill, M. M., Abrams, M. J. & Leuchter, A. F. Prospective testing of a neurophysiologic biomarker for treatment decisions in major depressive disorder: The PRISE-MD trial. *Journal of Psychiatric Research* **124**, 159–165 (2020).
33. de la Salle, S., Jaworska, N., Blier, P., Smith, D. & Knott, V. Using prefrontal and midline right frontal EEG-derived theta cordance and depressive symptoms to predict the differential response or remission to antidepressant treatment in major depressive disorder. *Psychiatry Research: Neuroimaging* **302**, 111109 (2020).
34. Jaworska, N., de la Salle, S., Ibrahim, M.-H., Blier, P. & Knott, V. Leveraging Machine Learning Approaches for Predicting Antidepressant Treatment Response Using Electroencephalography (EEG) and Clinical Data. *Frontiers in Psychiatry* **9**, (2019).
35. Mumtaz, W., Xia, L., Mohd Yasin, M. A., Azhar Ali, S. S. & Malik, A. S. A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. *PLOS ONE* **12**, e0171409 (2017).
36. Rajpurkar, P. *et al.* Evaluation of a Machine Learning Model Based on Pretreatment Symptoms and Electroencephalographic Features to Predict Outcomes of Antidepressant Treatment in Adults With Depression. *JAMA Network Open* **3**, e206653 (2020).
37. Wu, W. *et al.* An electroencephalographic signature predicts antidepressant response in major depression. *Nature Biotechnology* **38**, 439–447 (2020).
38. Zhdanov, A. *et al.* Use of Machine Learning for Predicting Escitalopram Treatment Outcome From Electroencephalography Recordings in Adult Patients With Depression. *JAMA Network Open* **3**, e1918377 (2020).
39. Al-Kaysi, A. M. *et al.* Predicting tDCS treatment outcomes of patients with major depressive disorder using automated EEG classification. *Journal of Affective Disorders* **208**, 597–603 (2017).
40. Zandvakili, A. *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. *Journal of Affective Disorders* **252**, 47–54 (2019).
41. Zhang, A., Yang, B. & Huang, L. Feature Extraction of EEG Signals Using Power Spectral Entropy. in *2008 International Conference on BioMedical Engineering and Informatics* 435–439 (IEEE, 2008). doi:10.1109/BMEI.2008.254.
42. Mumtaz, W., Ali, S. S. A., Yasin, M. A. M. & Malik, A. S. A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD). *Medical & Biological Engineering & Computing* **56**, 233–246 (2018).

43. Mumtaz, W., Xia, L., Mohd Yasin, M. A., Azhar Ali, S. S. & Malik, A. S. A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. *PLOS ONE* **12**, e0171409 (2017).
44. Cao, Z. *et al.* Identifying Ketamine Responses in Treatment-Resistant Depression Using a Wearable Forehead EEG. *IEEE Transactions on Biomedical Engineering* **66**, 1668–1679 (2019).
45. de la Salle, S., Jaworska, N., Blier, P., Smith, D. & Knott, V. Using prefrontal and midline right frontal EEG-derived theta cordance and depressive symptoms to predict the differential response or remission to antidepressant treatment in major depressive disorder. *Psychiatry Research - Neuroimaging* **302**, (2020).
46. Zhdanov, A. *et al.* Use of Machine Learning for Predicting Escitalopram Treatment Outcome From Electroencephalography Recordings in Adult Patients With Depression. *JAMA Netw Open* (2020) doi:10.1001/jamanetworkopen.2019.18377.
47. Deeks, J. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* **323**, 487–487 (2001).
48. Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J. & Bossuyt, P. M. M. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* **56**, 1129–1135 (2003).
49. Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G. M., de Bruin, H. & MacCrimmon, D. J. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. *Clinical Neurophysiology* **124**, 1975–1985 (2013).
50. Rabinoff, M., Kitchen, C. M. R., Cook, I. A. & Leuchter, A. F. Evaluation of Quantitative EEG by Classification and Regression Trees to Characterize Responders to Antidepressant and Placebo Treatment. *The Open Medical Informatics Journal* **5**, 1–8 (2011).
51. Sadat Shahabi, M., Shalbaf, A. & Maghsoudi, A. Prediction of drug response in major depressive disorder using ensemble of transfer learning with convolutional neural network based on EEG. *Biocybernetics and Biomedical Engineering* **41**, 946–959 (2021).
52. Howes, O. D., Thase, M. E. & Pillinger, T. Treatment resistance in psychiatry: state of the art and new directions. *Molecular Psychiatry* **27**, 58–72 (2022).
53. Jaworska, N., de La Salle, S., Ibrahim, M. H., Blier, P. & Knott, V. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. *Frontiers in Psychiatry* (2019) doi:10.3389/fpsyt.2018.00768.
54. Noble, W. S. What is a support vector machine? *Nature Biotechnology* **24**, 1565–1567 (2006).
55. Lin, C. C. *et al.* Evaluation of a Machine Learning Model Based on Pretreatment Symptoms and Electroencephalographic Features to Predict Outcomes of Antidepressant Treatment in Adults With Depression: A Prespecified Secondary Analysis of a Randomized Clinical Trial. *Scientific Reports* **7**, 1–8 (2020).
56. Cao, B. *et al.* Predicting individual responses to the electroconvulsive therapy with hippocampal subfield volumes in major depression disorder. *Scientific Reports* **8**, 1–8 (2018).
57. Jatoi, M. A., Kamel, N., Malik, A. S., Faye, I. & Begum, T. A survey of methods used for source localization using EEG signals. *Biomedical Signal Processing and Control* **11**, 42–52 (2014).
58. Lachaux, J. P., Rudrauf, D. & Kahane, P. Intracranial EEG and human brain mapping. *Journal of Physiology-Paris* **97**, 613–628 (2003).
59. Naimi, A. I. & Balzer, L. B. Stacked generalization: an introduction to super learning. *European Journal of Epidemiology* **33**, 459–464 (2018).
60. Sesmero, M. P., Ledezma, A. I. & Sanchis, A. Generating ensembles of heterogeneous classifiers using Stacked Generalization. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **5**, 21–34 (2015).
61. Claesen, M. & de Moor, B. Hyperparameter Search in Machine Learning. 10–14 (2015).
62. Bardenet, R., Brendel, M., Kégl, B., Sebag, M. & Fr, S. *Collaborative hyperparameter tuning*. vol. 28 (2013).
63. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. (2012).

Chapter 7 - Data-driven biomarkers of treatment response within randomized clinical trials in psychiatry: A systematic review and methodological recommendations for machine-learning precision trials

Authors: Devon Watts MSc[1]; Diego Librenza-Garcia, MD, PhD[2,3]; Pedro Ballester MSc[1]; Bruno Jaskulski Kotzian[2]; Jessica Yang[4]; Benício Frey, MD, PhD[1,5]; Luciano Minuzzi, MD, PhD[1,5]; Benson Mwangi, PhD[6]; Flávio Kapczinski, MSc, MD, PhD, FRCPC[1,2,5]; Ives Cavalcante Passos, MD, PhD[2,3]

1. Neuroscience Graduate Program, McMaster University, Hamilton, Canada

2- Laboratory of Molecular Psychiatry, Centro de Pesquisa Experimental (CPE) and Centro de Pesquisa Clínica (CPC), Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, RS, Brazil. Instituto Nacional de Ciência e Tecnologia Translacional em Medicina (INCT-TM), Porto Alegre, RS, Brazil.

3- Universidade Federal do Rio Grande do Sul, School of Medicine, Graduate Program in Psychiatry and Behavioral Sciences, Department of Psychiatry, Porto Alegre, RS, Brazil.

4. College of Pharmacy, University of Texas at Austin, Austin, Texas, USA

5. Department of Psychiatry and Behavioral Neurosciences, McMaster University, Hamilton, ON, Canada.

6. Department of Psychiatry and Behavioral Sciences, The University of Texas Science Center at Houston, Houston, Texas, USA

*Corresponding author:

Ives Cavalcante Passos, MD, PhD

Professor of Psychiatry

Federal University of Rio Grande do Sul, Avenida Ramiro Barcelos, 2350, Zip Code: 90035-903, Porto Alegre-RS, Brazil, Phone: +55 512 101 8845, Email: ivescp1@gmail.com

Email for all authors:

Devon Watts: wattsd@mcmaster.ca

Diego Librenza-Garcia: librenzagarcia@gmail.com

Pedro Ballester: pedballester@gmail.com

Bruno Jaskulski Kotzian: brunokotzian@hotmail.com

Jessica Yang: jessica.yang10@gmail.com

Benício Frey: freybn@mcmaster.ca

Luciano Minuzzi: minuzzi@mcmaster.ca

Benson Mwangi: benson.mwangi@gmail.com

Flávio Kapczinski: flavio.kapczinski@gmail.com

Ives Cavalcante Passos: ivescp1@gmail.com

Word Count: 9561

This chapter in its entirety is currently under ***revision*** in the journal Molecular Psychiatry.

ABSTRACT

Background: Selecting a psychotropic medication in psychiatry remains a trial-and-error process, with no specific biomarker to lend support in clinical decision making. Additionally, randomized clinical trials (RCTs) and meta-analyses yield group-level results, and usually do not adequately model the heterogeneity and multimorbidity observed in patients with psychiatric disorders. There is, therefore, a critical need for predictive tools to aid clinicians in determining the likelihood an individual patient will respond to a given treatment.

Aims: The aims of the present study were 1) to review machine learning models of treatment response within clinical trials in psychiatry that incorporate data-driven biomarkers and 2) to provide methodological recommendations for machine-learning precision trials, a new trial design to occur following the successful completion of an RCT.

Method: We performed a systematic review of studies using data from randomized clinical trials to predict treatment response in patients with psychiatric disorders using machine learning models comprising biological or physiological input features (Registration Number: CRD42016049635) by searching PubMed, Scopus, and Web of Science for articles published between January 1981 and March 2022.

Results: We included 26 studies that predicted treatment response using data from randomized clinical trials among patients with any psychiatric diagnosis (n = 7031 patients in total). Studies thus far have used resting-state and task-specific fMRI, resting-state EEG, structural MRI, blood metabolites, serum biomarkers, and single nucleotide polymorphisms to develop predictive models. Model performance within classification models ranged from 57-86.7% when using

peripheral blood markers, 76-81% when using EEG, 71.4-77.5% when using neuroimaging features, and 50.3-84% when using multimodal data, respectively. Furthermore, based on the consistency of performance across models with large sample sizes, the highest degree of evidence was in predicting response to sertraline and citalopram using fMRI features. However, prospective models with larger sample sizes using EEG, blood-biomarker and multimodal data are required to determine whether a specific modality is superior in predicting treatment response.

Conclusions: Machine learning models with high-quality input variables have the potential to address some limitations of evidence-based medicine, shifting the focus from group-level results to individualized predictions. We present methodological recommendations for machine-learning precision trials, an important second step following RCTs to improve the generalizability of models to heterogeneous patients seen in the clinic. Moreover, machine-learning precision trials of treatment selection, evaluating individual differences in comparative effectiveness across the same group of patients, are needed to advance the field of precision psychiatry.

Keywords: machine learning; predictive modelling; artificial intelligence; precision psychiatry; computational neuroscience; biomarkers; evidence-based medicine, treatment response

INTRODUCTION

Evidence-based medicine (EBM) has prompted a revolution in patient treatment since its introduction in research and clinical practice. Indeed, EBM has led to several improvements in methodological research standards, as well as clinical guidelines and knowledge translation [1]. The gold-standard of EBM are randomized controlled trials (RCTs), which assess the average

group response to a given intervention [2]. Moreover, meta-analyses of RCTs, which pool individual RCTs together to derive an overall estimate of the effect of the intervention, are a key component of EBM [3].

Considering that individual patients may deviate from the average group response, it can be expected that a specific treatment with demonstrated efficacy, relative to placebo, may not be efficacious across all patients. Additionally, due to strict inclusion/exclusion criteria meta-analyses and RCTs cannot properly map the complexity that are often seen in real patients, and as a result, are unable to render tailor-made evidence [4]. In fact, the very idiosyncrasies that characterise most patients, such as multimorbidity profiles, are often exclusion criteria in clinical trials.

It is also important to mention that statistically significant associations at the aggregate level do not necessarily translate into clinical benefit. For instance, in a network meta-analysis comparing the efficacy and acceptability of 21 antidepressant drugs across 522 trials for the acute treatment of adults with Major Depressive Disorder (MDD), while all antidepressants were found to be more efficacious than placebo, significant variability in efficacy and acceptability was observed between medications in head-to-head trials [5]. Similar heterogeneity in treatment efficacy was also observed across patients with schizophrenia in a network meta-analysis comprising 402 trials and 32 oral antipsychotics, with large differences in side effects between medications [6]. Altogether, available evidence suggests that approximately 20-60% of patients with psychiatric disorders continue to show significant residual symptoms following a course of treatment of sufficient dose and duration [7].

Despite clinical heterogeneity in response to medications that have been shown to be effective in randomized placebo-controlled trials, we currently lack objective biomarkers to guide the clinical

likelihood of sufficient symptomatic improvement, inadequate symptom reduction, or remission within a specific patient to a given course of treatment. As such, patients continue to endure prolonged periods of "trial-and-error" in search of effective treatment and the burden associated with this process. Moreover, validated, and reliable biomarkers are needed to improve our understanding of the mechanisms of patient remission in response to specific treatments. For instance, while first-line antidepressants such as fluoxetine have been shown to be effective in many patients with depression for over 3 decades [8], debate remains surrounding their exact mechanisms of action [9]. Therefore, new strategies are required to determine which treatments are likely to be effective for a given patient, expedite biomarker discovery, and improve our mechanistic understanding of how currently approved medications improve symptoms, to guide the development of next-generation therapeutics in psychiatry.

Towards this end, machine learning is a subfield of artificial intelligence focused on computational methods that can extract relevant information from complex datasets [10]. Such methods can model patterns to generate individualized predictions using high quality data from various modalities, such as neuroimaging, genetics, neurophysiology, and clinical features [11]. Incorporating these techniques into less restricted clinical trials with medications that have already proven their efficacy in previous RCTs will aid in the development of precision psychiatry, by enabling more precise interventions that include patient's idiosyncrasies [12]. Considering the limitations of a "trial-and-error" approach to treatment in psychiatry, there is a major unmet need for individualized predictions of response to treatment.

In the present study, we aimed to systematically review studies that used machine learning techniques to predict treatment response within randomized clinical trials in patients with psychiatric disorders. To assess predictors that may be implicated in the underlying mechanisms

of action of treatment response, only studies that incorporated biomarkers, broadly defined as biological or physiological input features, were included. Additionally, we provide recommendations for a new trial design that should be conducted following successful RCTs. We refer to this as *machine-learning precision trials.*

METHODS

This study has been registered on PROSPERO with the registration number PROSPERO CRD42019127169.

Search strategy

Three electronic databases (PubMed, Scopus, and Web of Science) were examined for articles published between January 1981 and March 2022. To identify relevant studies, the following structure for the search terms was used: (Artificial Intelligence OR Supervised Machine Learning) AND (psychiatric disorders) AND (clinical trials OR treatment response OR treatment prediction OR treatment selection). The complete filter is available in the supplementary material. We also screened the references from the articles included to find potential missed articles. There were no language restrictions.

Eligibility criteria

This systematic review was performed according to the PRISMA statement [13]. We selected original articles that assessed patients with a psychiatric disorder treated with pharmacological or non-pharmacological interventions coupled with machine learning models to predict treatment outcomes. Review articles, observational studies, naturalistic trials, non-interventional studies,

models predicting response to heterogeneous open-label treatments (e.g., several SSRI antidepressants), and studies which did not consider biological or physiological variables as candidate features were excluded. Furthermore, studies that lacked either cross-validation measures or training and testing sets were excluded. Additionally, non-randomized open-label trials are considered separately in the supplementary material, as they lack a comparator group to assess the specificity of the predictive model to the treatment of interest, relative to a placebo or other treatment arm.

Data collection and extraction

Initially, the potential articles were independently screened for title and abstract contents by two researchers (DW and DLG). Then, they also obtained and read the full text of potential articles. A third author (ICP) provided a final decision in cases of disagreement. Data extracted from the studies included publication year, sample size, diagnosis, data inputted into the machine learning model, machine learning algorithm, sampling method and data imputation, type of intervention, outcomes of interest, and statistical performance of the models (i.e., accuracy, balanced accuracy, sensitivity, specificity, area under the curve, true positive, false positive, true negative and false negative and confidence intervals of performance metrics, when available). We developed a quality assessment instrument specific to machine learning studies since there is no tool for quality assessment in machine learning studies. This instrument is further described in the supplementary material.

RESULTS

We found 16,669 potential abstracts and included 26 articles in the present review, three included after reference screening [14–16]. A list of included studies, comprising the clinical sample,

outcome, machine-learning models, top data-driven biomarkers, and performance metrics, can be observed in Table 1. Furthermore, Table 2 provides a thorough overview of true and false positives and negatives across classification studies, methods to address class imbalance as well as 95% confidence intervals of model accuracy. Additionally, details related to data pre-processing and feature extraction for each model can be found in Supplementary Table S3. A quality assessment developed for machine learning models can be observed in Supplementary Table S2. Studies with lower quality assessment scores are described in brevity in the results section. Additionally, further context regarding feature extraction and model development are provided in the results, where required. Of the included studies, 3 studies used peripheral blood markers [17–19], 5 studies used electroencephalography [16,20–23], 9 studies used neuroimaging [24–32], and 10 studies used multimodal data [15,26,33–40], defined as at least two feature modalities, or overarching categories of input features, such as fMRI and EEG predictors. Furthermore, a table containing studies that developed models using data from non-randomized open-label trials can be found in Supplementary Table S1. Among them, 6 studies used EEG, 13 studies used neuroimaging, and 6 studies used multimodal data.

Studies using blood biomarkers and genetics

Three studies developed predictive models of treatment response within randomized clinical trials using peripheral blood markers [17–19]. Amminger and colleagues predicted response to omega-3 fatty acids (ω-3) or placebo in a 12-week randomized controlled trial (RCT) of 81 individuals at ultra-high risk of psychosis. Fatty acid composition was quantified via capillary gas chromatography. Clinical response was defined as a ≥15-point increase in Global Assessment of Functioning (GAF) scores from baseline to the end of treatment. Erythrocyte fatty acids were used as predictive variables, comprising six categories of fatty acids, including

arachidonic acid (AA), eicosapentaenoic acid (EPA), and docosahexaenoic acid (DHA). Using a Gaussian Process Classifier, their model showed an accuracy of 86.7% in predicting response to Polyunsaturated fatty acids (PUFAs), and 79.6% in predicting response to placebo. Important variables in the ω-3 model included nervonic acid, margaric acid, and arachidonic acid [17].

Furthermore, Maciukiewicz et al. [19] used SNP data from three previously conducted RCTs of duloxetine or placebo for 8 weeks, to predict treatment response and remission, defined as a >50% change in the Montgomery-Asberg Depression Rating Scale (MADRS) from baseline and a total MADRS score ≤10 at endpoint, respectively. However, the model showed poor balanced accuracy (46-49%), defined as the arithmetic mean of sensitivity and specificity [41], was observed across models. Additionally, Hou et al. [18]predicted response to 11-weeks of ondansetron, a 5-HT3 receptor antagonist, vs. placebo in 251 patients with Alcohol-Use Disorder (AUD) using polymorphisms in the promoter region of the SLC6A4 gene. However, the accuracy of these models was not reported [18].

Studies using Electroencephalography

Five studies developed predictive models of treatment response within RCTs using pre-treatment EEG [16,20–22,42]. Wu and colleagues developed a predictive model of response to sertraline or placebo using data from 228 patients with MDD enrolled in the Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care for Depression (EMBARC) trial. Treatment response was considered as a continuous outcome, using pre- minus post-treatment differences in the 17-item Hamilton Depression Rating Scale (HAMD-17). The authors also developed a predictive algorithm known as Sparse EEG Latent SpacE Regression (SELSER) which uses spatial filters that map EEG signals to a latent space performed under a sparse

constraint on the number of spatial filters, and then relates the band powers of the latent signals to a treatment outcome using a linear regression model. In a leave-study-site-out analysis, SELSER predicted response to sertraline with a Pearson's r (r) of 0.60 and predicted response to placebo with r=0.41. Of note, when models were applied to the opposite arm of the study, the outcome could not be predicted (r=-0.03 and r=<0.22, respectively), demonstrating their specificity. Moreover, for the sertraline arm, only signals from the resting-eyes open condition were significantly predictive of treatment score change during cross-validation [23].

Additionally, Cao and colleagues predicted treatment response (≥ 45% reduction in Hamilton Depression Rating Scale 17-items ($HDRS_{17}$) from baseline to 240 min post-treatment) in a double-blind placebo-controlled trial of ketamine (0.2mg/kg, 0.5 mg/kg) and saline. Using EEG power and alpha asymmetry features, their Support Vector Machine (SVM) model showed an accuracy of 78.4% [21]. Furthermore, de la Salle and colleagues predicted clinical response (≥ 50% improvement in MADRS scores from baseline) within a 12-week trial of bupropion, escitalopram, or combined treatments, across 47 patients with treatment resistant depression. Within a logistic regression model, prefrontal cordance across delta, theta, alpha, and beta frequency bands and change scores in middle right frontal cordance resulted in an accuracy of 81% and 74% in predicting clinical response, respectively. Similarly, clinical remission could be predicted with 70% accuracy using prefrontal cordance, however middle right frontal cordance features were not discriminative (51% accuracy) [16]. In another study, using the same dataset, Jaworska and colleagues predicted response in a double-blind trial of escitalopram + bupropion, escitalopram + placebo, or bupropion + placebo, resulting in an AUC of 0.716-0.901 [22]. Furthermore, another study [20] predicted response to transcranial direct current stimulation (tDCS)

in a sample of 10 patients with MDD, resulting in a cross-validated accuracy of 76%, with the best performance using FC4-AF8 electrode pairs [20].

Studies using Neuroimaging

Nine studies [24–32] developed predictive models of treatment response within randomized clinical trials using neuroimaging derived features. Braund and colleagues [24] developed a model using a connectome signature associated with neuroticism and clinical features to predict treatment response, defined as >50% reduction in $HDRS_{17}$. More specifically, baseline intrinsic functional connectivity was calculated between each pair of 400 cortical regions and 36 subcortical regions, analyzed using network-based statistics to identify connectomics features associated with neuroticism (total NEO-FFI scores). This network-based statistics (NBS) analysis identified a signature comprising 622 connections across 198 nodes, where greater neuroticism was associated with significantly higher functional connectivity (corrected p=.010). Using an SVM model, with a filter-based feature selection method, 19 connections across 30 brain regions correctly classified responders from non-responders with 75% accuracy [24].

Fonzo and colleagues [26] developed a model using emotional conflict-regulation-related brain activity, to a previously characterized emotional conflict task as part of EMBARC, the largest neuroimaging-coupled placebo-controlled RCT of depression to date. In total, 309 medication-free outpatients with depression received either the SSRI sertraline or placebo for 8 weeks. Following a series of pre-processing steps, as outlined in Supplementary Table S3, regions of interest (ROIs) were mapped to seven functional networks according to the spatial overlap between each ROI and each network. In a regression model, treatment response was defined as pre-minus post-treatment change in the HAMD-17 using emotional conflict regulation activation. A relevance vector machine (RVM) model trained on the sertraline outcome yielded a

cross-validated prediction of r=-0.49, and when applied to the placebo arm, the sertraline-trained model did not yield a significant prediction of HAMD-17 change (r=-0.06). Interestingly, an RVM model trained on emotion conflict regulation brain activation data in the placebo arm to predict placebo outcome did not yield significant correlations between model-predicted symptom changes and observed symptom changes in either the placebo or sertraline arms (r=0.11, P>0.20). This suggests that the model reflects a sertraline-specific signal separate from treatment effects present across both treatment arms. Important features specific to the sertraline RVM model included the right insular lobe and right middle temporal gyrus [26].

Koutsouleris et al. [28] predicted response to repetitive transcranial magnetic stimulation (rTMS) for first-episode psychosis within a multi-site trial of 92 patients randomized to either active (N=45) or sham (N=47) 10-Hz rTMS applied to the left dorsolateral prefrontal cortex (DLPFC) over 15 sessions. Response and nonresponse were defined according to a ≥ 20% change in PANSS negative scores through treatment. Features were extracted from structural MRI data using their NeuroMiner tool (https://github.com/neurominer-git/). Using a linear SVM, they correctly separated PANSS responders from non-responders with a cross-validated balanced accuracy of 84.3%. Important features included relative gray matter density (GMD) reductions in prefrontal, insular, and medial cortices. Of note, this pattern specifically separated nonresponders from responders in the active, but not the sham treatment group [28].

Furthermore, Nord and colleagues. [30] conducted an 8-session double-blind RCT of real (N=20) or sham (N=19) transcranial direct current stimulation (tDCS) as an adjunct to cognitive behavioral therapy (CBT) in 39 unmedicated patients with MDD. An MRI protocol involved a T1-weighted anatomical scan and two T2-weighted functional scans during the n-back working memory task and an emotional processing task, where participants discerned the gender of

fearful, happy, and neutral faces. Immediately prior to each CBT session, a 1 mA constant current was delivered to the left PFC (anode on F3) using an EEG cap for placement, and a cathode on the ipsilateral deltoid. Treatment response was defined according to a ≥50% reduction in the HAM-D from baseline to end of treatment. Patients were divided according to low and high L-DLPFC activation during the working memory task, and baseline L-DLPFC activation was shown to discriminate responders from non-responders with an Area Under the Curve (AUC) of 0.856. Of note, this same pattern of activation did not discriminate responders from non-responders in the sham condition (AUC=0.417) [30].

Sarpal and colleagues [31] predicted response within a double-blind randomized controlled trial of either risperidone or aripiprazole for 12 weeks in 81 patients with first-episode schizophrenia. Patients underwent a resting-state fMRI scan, and 91 features were extracted using a striatal connectivity index calculation, which comprised functional connectivity in the striatum. Treatment response was defined as two consecutive visits with a Clinical Global Impression (CGI) improvement score of 1 or 2 (very much improved) and a rating of 3 (mild) or less in conceptual disorganization, grandiosity, hallucinatory behaviour, and unusual thought content on the Brief Psychiatric Rating Scale (BPRS). Using cox regression, their model showed an accuracy of 77.5%. In posterior regions, greater connectivity in striatal subdivisions at baseline was associated with better subsequent treatment response. Conversely, lower striatal connectivity of frontal nodes at baseline was associated with better subsequent response [31].

Furthermore, Yip and colleagues [32] developed a model to predict abstinence within a 12-week randomized controlled trial of behavioural therapy plus galantamine or placebo for cocaine use disorder. Abstinence during treatment was determined using results of biweekly urine testing and defined as the percentage of cocaine-negative urine provided during treatment. fMRI data was

acquired during the performance of a monetary delay task, and Connectome-Based Predictive Modelling (CPM) was used to extract features, which included group connectivity matrices as input to generate a predictive model of the outcome of interest from connectivity matrices. Briefly, edges and behavioural data from the training set are correlated using regression analyses to identify positive and negative predictive networks. Positive networks comprise those where increased edge weights (increased connectivity) are associated with the variable of interest, and negative networks are those where decreased edge weights (decreased connectivity) are associated with the variable of interest. In an independent sample, abstinence during treatment was predicted with r=0.36 (p=0.016), and with 64% accuracy when dichotomizing patients into the presence or absence of any cocaine-negative urine. The highest-degree nodes (those with the most connections) in the positive network were characterized by more within-network connections across medial, frontal, frontoparietal, default mode, motor/sensory, visual association, and salience networks. Within negative networks, more connections were observed within occipital and subcortical networks [32].

Additionally, Nemati and colleagues [29] identified a specific connectome fingerprint that predates and predicts response to monoaminergic antidepressants. Data used in the predictive model involved 202 individuals with MDD from the EMBARC trial, and 56 individuals with MDD from a previous RCT of ketamine, from which baseline fMRI data was available. Features were extracted from fMRI data using nodal internal network restricted strength (niRNS), calculated as the average connectivity between nodes and all other nodes within the same intrinsic connectivity network (ICN), and nodal external network restricted strength (neNRS), calculated as the average connectivity between each node and all other nodes outside its ICN, respectively. Brain nodes were defined using multimodal parcellation atlases, dividing the cerebral cortex,

subcortical regions, and cerebellum into 424 nodes. The full connectome was calculated as the pairwise correlation coefficient between the averaged time series, which was subsequently transformed using a Fisher's z transformation. Additionally, a network restricted strength predictive model (NRS-PM) was used, which incorporates feature selection to identify NRS edges that positively or negatively predict the behavioural measure of interest ($p < 0.05$), and following this, the weighted sum of positive edges minus the weighted sum of negative edges are used to generate a summary statistic for each subject, with the resulting coefficients applied to predict the outcome. The whole brain NRS-PM predicted antidepressant response across AA-4 and AA-150 architectures, following false discovery rate (FDR) correction, with a peak at AA-58 ($r=0.27$, CV=10, iterations=1000, $p=0.003$). Independently, the positive predictive edges peaked at AA-58 ($r=0.29$, CV=10, iterations=1000, $p=0.001$) and the negative predictive edges peaked at AA-26 ($r=0.25$, CV=10, iterations=1000, $p=0.003$). Interestingly, the model showed partial generalization to an independent ketamine dataset, where it predicted response to ketamine compared with lanicemine ($r=0.55$, $p=0.0003$), but not ketamine relative to placebo [29].

Furthermore, one study predicted response to sertraline using the EMBARC trial but did not find differences in features between sertraline and placebo models [25], and another study [27] predicted response to citalopram or placebo using a network-based statistical analysis, resulting in an AUC of 0.68 in predicting response, and an AUC of 0.73 in predicting remission, respectively.

Studies using multimodal predictors

Ten studies [15,26,33–40] predicted treatment response within randomized clinical trials using multimodal data. Crane and colleagues [34] developed a predictive model of treatment response to antidepressants using inhibitory control during a functional MRI. Twenty-nine patients with

MDD, free of any antidepressants for at least 90 days, were treated with open-label escitalopram or duloxetine for 10 weeks. The parametric go/no-go test (PGNG), which measures attention, set-shifting, processing speed and correct/incorrect responses was used, and ICA beta weights within the PGNG imaging task, traditional haemodynamic response function (HRF), medication type, age, sex, and the interaction between component beta weights and medication group were used as predictors. Following leave-one-out cross-validation, a random forest model predicted treatment remission (post-treatment $HDRS_{17}$ <8) with 84% accuracy [34].

Furthermore, Taliaz and colleagues [40] developed a predictive model of treatment response to antidepressants using 1697 patients with MDD from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial, which was tested on an external sample of 132 patients treated with citalopram from the Pharmacogenomic Research Network Antidepressant Medication Pharmacogenetic Study (PGRN-AMPS). Two measures of treatment response were used, corresponding to classic response and exponential response. Classic response was defined as a ≥50% reduction in the Quick Inventory of Depressive Symptomatology (QIDS) from baseline. Conversely, exponential response involved a continuous measure that represented the exponential fit for the individual longitudinal measurement of QIDS, during a specific treatment. This measure accounts for the change of the score over time, as well as the speed and dynamics of the response. As such, the median of the exponential antidepressant improvement rates was calculated independently for each STAR*D treatment. Candidate features for the models included clinical and demographic variables, alongside genes and microRNA that were reported to be associated with depression, antidepressant response, metabolism, and side effects, yielding 281 genetic components. In the validation set, a SVM model predicted response with 72.3% across medications, with the best performance in predicting response to venlafaxine, with a

balanced accuracy of 80.2%. Furthermore, when tested on the external PGRN-AMPS dataset, the citalopram model showed a balanced accuracy of 61.3% [40].

Fonzo and colleagues [14] predicted treatment response within an RCT comprising 12 sessions of prolonged exposure treatment (N=36) or waitlist condition (N=30) in patients with Post-Traumatic Stress Disorder (PTSD). Clinical remission was defined as a post-treatment Clinician-Administered PTSD Scale (CAPS) score ≤ 20. Using a combination of baseline clinical features, treatment arm, and bilateral activation of several brain regions during an emotional reactivity task, clinical remission was predicted with an accuracy ranging from 79.5%-97.7%. However, it should be noted that model accuracy is likely higher than what would be expected when tested in an independent cohort, since the same sample of participants was used in training and testing its predictive accuracy. Nevertheless, the most important features in predicting total CAPS scores from linear mixed models included 23 regions during an emotional reactivity task, 7 regions during an emotional conflict task, and 1 region in an emotional conflict vs gender conflict task. A summary of these predictors can be observed in Table 1 [14].

Joyce and colleagues [36] developed a predictive model of treatment response using data from PGRN-AMPS, an 8-week clinical trial of escitalopram or citalopram comprising 529 patients with MDD. Model performance was tested on CO-MED, a 7-month clinical trial where patients were randomized to either escitalopram + placebo, bupropion + escitalopram, or extended-release venlafaxine plus mirtazapine. Clinical response was defined as ≥50% reduction in QIDS-C total score from baseline, and remission was defined as a score of ≤5 on the QIDS-C, respectively. Two predictive models were developed, one comprising clinical, sociodemographic, and metabolomic features common to both the PGRN-AMPS and CO-MED studies, and a second augmented model incorporating six previously functionally validated

SNPs. Using the "metabolomics models" feature set, the best trained classifiers predicted response to combination antidepressant therapies at 8 weeks with accuracies of 76.6% ($p<0.005$; AUC:0.85) and 72.7% ($p=0.053$; AUC:0.76) for penalized regression and XGBoost, respectively. Using the "multi-omics models" feature set, accuracies improved to 77.5% ($p<0.01$; AUC:0.86) and 76.1% ($p=0.017$; AUC: 0.83). Of note, performance slightly decreased in the SSRI only XGBoost model when combining metabolomic and SNP data, relative to metabolomic data alone (75.3% vs 73.2%) [36].

Furthermore, Nguyen and colleagues [37]predicted change scores in $HAMD_{17}$, clinical response (≥50% reduction in $HAMD_{17}$ at week 8), and clinical remission (≤7 HAMD17 at week 8), using data from the EMBARC study comprising 222 patients randomized to 9 weeks of sertraline (n=106) or placebo (n=116). Subsequently, sertraline non-responders (n=37) were switched to 8 weeks of bupropion. Reward task-based fMRI was acquired at baseline visit for 8 minutes during a block-design number-guessing task, where participants' differential brain activation was measured between punishing vs. rewarding trials. Contrast maps were quantified using brain activation in the anticipation phase of number-guessing trials, reward expectancy, and prediction error, and were parcellated into 200 functional brain regions, yielding 600 fMRI features for each patient. Additionally, 95 pre-treatment clinical and demographic features were also considered as candidate features in feed-forward neural network models. Change scores in $HAMD_{17}$ were predicted in sertraline, placebo, and bupropion conditions, with an R^2 of 0.48, 0.28, and 0.34, respectively [37]. Additional performance metrics can be observed in Table 1.

In another study, Rajpurkar and colleagues [38] predicted improvements in depressive symptoms within a clinical trial of escitalopram, sertraline, or extended-release venlafaxine, using resting-state EEG and baseline $HDRS_{17}$ scores. EEG features included absolute power, relative power,

frontal alpha asymmetry, and beta-alpha ratio. In the combined model, the authors reported an R^2 of 0.551. Symptom features alone resulted in an R^2 of 0.375. Of note, Shapley Additive Explanations (SHAP) were used to quantify the effect of each feature on the models. SHAP values were aggregated for features on individual predictions and the averaged Shapley contributions were reported as a percentage of the associations of all features. Shapley contributions were reported for changes in each item of the $HDRS_{17}$ including waking early, physical anxiety, and trouble sleeping. For each individual item, baseline HRSD-21 scores showed the highest contribution. For instance, baseline trouble sleeping showed a 57.3% contribution to changes in trouble sleeping throughout treatment, followed by T7-T3 alpha absolute ratio (6.7%) and T7-T3 beta absolute ratio (4.4%), respectively [38]. Further details on important EEG features within the model can be observed in Table 1.

Other study [35] predicted response within the NIMH-funded Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) [43] study, resulting in an accuracy between 55-66%, a comparative trial of several antipsychotics [33], resulting in an accuracy of 50.3%, and an RCT of exercise therapy [39] in patients with MDD, resulting in an AUC of 0.785 in predicting remission, and an AUC of 0.710 in predicting non-response, respectively. Furthermore, Athreya and colleagues [15] predicted remission and response in patients with MDD, comprising 398 patients from PGRN-AMPS, 467 patients from STAR*D, and 165 from the International SSRI Pharmacogenomics Consortium (ISPC) trials. Using plasma metabolites associated with SSRI response and SNPs, remission and response was predicted with an AUC of ~0.70 across trials [15].

DISCUSSION

Within this review, several studies have developed predictive models of treatment response using biological and physiological features generated from well-characterized, large-scale placebo-controlled trials, which allows for a comparative examination of data-driven biomarkers that are specific to the active treatment arm. Importantly, a subset of these models [23,33,36,40] were replicated in independent datasets, largely maintaining meaningful but modest predictive accuracy, which suggests the potential for their scalability as classification tools.

Model Performance

In terms of performance, models using peripheral blood markers ranged from 66-86.7% accuracy in predicting response across two studies [17,19], whereas a third study [18] only reported mean difference of the continuous outcome identified between algorithms. EEG models ranged from 76-81% accuracy to predict response across three studies [20–22], whereas a third regression model yielded an R^2 of 0.60 in predicting response to sertraline, relative to placebo (R^2=0.41) [23]. A larger proportion of studies thus far have used neuroimaging, where the vast majority have involved features extracted using fMRI. Roughly half (44.4%) of all neuroimaging models [25,26,29,32] have predicted continuous outcomes, with an R^2 ranging from 0.19-0.49 in predicting response to sertraline (median = 0.346, n=3), and similar performance in a model to predict cocaine abstinence (R^2= 0.36). In four studies using classification-based outcomes that reported model accuracy,[24,28,31,32], performance ranged from 71.4%-77.5%, with the best performance in a small trial (n=41) of patients within an RCT of risperidone or aripiprazole [31]. Within multimodal models, comprising six studies, accuracies ranged from 50.3-84%, with the best performance in a small trial of 49 medication-free patients in predicting response to open-label escitalopram or duloxetine for 10 weeks [34].

Model Validation

However, it is important to mention that a subset of studies [20,26,27,30–32,34,37], especially those with smaller sample sizes, did not incorporate a holdout set, characterized as a partition of the sample that the model is tested on following cross-validation [44]. Evaluating model performance using standard cross-validation alone may lead to inflated metrics, as discussed elsewhere [45]. As such, when a sample is of sufficient size to train a model, keeping approximately 30% of the data as a holdout set to test model performance is useful to assess overfitting, and provide a more realistic appraisal of model generalizability [46].

Moreover, best practices would involve testing model performance on an external sample, such as another RCT containing the same input features and outcome, to provide further insights into the generalizability of the model to other datasets evaluating the same intervention. Six studies [19,21,24,28,36,40] included in the present review tested the performance of their models on independent samples. For instance, Taliaz and colleagues [40] utilized a subset of the STAR*D dataset where genetic variables were available, comprising 1697 patients, separated into training (n=1167), testing (n=271), and internal validation sets (n=259). Additionally, they assessed performance on an external validation set (n=132) from the PGRN-AMPS study, which largely preserved model performance (67% in internal validation set vs 61% in external validation set) [40].

Furthermore, testing model accuracy on external datasets also provides an opportunity to assess the specificity of the model, by determining whether its performance generalises to different treatment arms and evaluating mutually exclusive and overlapping features. For instance, Nemati and colleagues [29] tested whether their sertraline model generalised to a dataset from an independent ketamine RCT. Their model was found to predict response to ketamine compared

with lanicemine (r=0.55, p=0.0003), but not ketamine relative to placebo [29]. As such, this approach allows for a comparative assessment of treatment specific biomarkers of response, relative to placebo, and other interventions, which may help inform treatment-specific mechanisms of therapeutic efficacy.

The following sections discuss data-driven biomarkers of treatment response identified in randomized clinical trials, provides a comparison of data-driven biomarkers across RCTs, relative to randomized and non-randomized open-label trials, as well as considers common algorithms and pre-processing strategies, and quality-metrics across studies.

Data-driven biomarkers of treatment response in randomized and non-randomized clinical trials

Given the relatively small number of studies (n=26) included in the present review, data-driven biomarkers identified in randomized-controlled trials (n=18, 72%) and randomized open-label trials (n=7, 28%) were assessed relative to excluded non-randomized open-label trials (n=25, Supplementary Table S1) to identify the degree of consistency in data-driven biomarkers across studies. Considering only six studies thus far that have utilized SNPs from peripheral blood samples as predictive features of treatment response, the lack of overlap in top features across studies is unsurprising. Nonetheless, twenty-one SNPs and several genes of interest have been identified, including the protein encoding genes MTOR, TSPAN5, DEFB1, AHR, ERICH3, PRKCA, GRIA1, GRIN2A, IFNA1, FKBP5, and GRIK4, as further described in Table 1.

In terms of EEG studies, features in the alpha band were found to be highly predictive of treatment response in seven of eight studies (87.5%) where frequency band features were

included as candidate features. However, while higher alpha band power in prefrontal regions were found to be predictive of response to escitalopram [38], fluoxetine [47], and rTMS [48], lower alpha power in the Fp2 channel was found in responders in a small ketamine trial [21]. Furthermore, among four studies [48–51] that incorporated theta cordance as candidate features, it was found to be predictive of treatment response in 50% of cases [48,49].

Regarding neuroimaging studies using fMRI, BOLD signals in the amygdala [14,30], fusiform gyrus [52,53], and posterior cingulate cortex [31,54] were found to be predictive of treatment response across two studies. Additionally, three studies found activation in the DLPFC [30,32,54] and thalamus [31,54,55] to be predictive of treatment response, four found temporal gyrus activation to be predictive [26,32,52,56], and five studies found activation in anterior cingulate cortex to be among the top predictors of treatment response [27,31,34,54,57], respectively. Additionally, three studies (23,25,53) found increased pre-treatment functional connectivity in the default mode network to be predictive of treatment response.

Of note, features derived from fMRI that were found to be predictive of response to sertraline, relative to placebo, included activation in the right insular lobe and right middle temporal gyrus [14], positive predictive edges in AA-58, and negative predictive edges in AA-26 [29], as well as increased activation in the inferior frontal gyrus, pars triangularis [37].

Furthermore, among the five structural MRI studies, regions of the temporal lobe were found to be predictive across two studies [58,59], and structures found predictive in fMRI studies were also highlighted, including the insula [56,58] and cingulate cortex [58].

It is also important to highlight that many features identified within these treatment response prediction models are highly interrelated, and given this, important features may vary

dramatically across studies even in cases where the predictive accuracy of models are highly similar.

Top Algorithms and Pre-processing Strategies

Across 26 studies included in the present review, eight studies (30.7%) [18,20–22,33,36,39,60] assessed the relative performance of multiple algorithms. However, little consistency was observed in top algorithms across studies. For instance, while six studies included SVM, it was found to be the best performing algorithm in only one case [21]. Similarly, three studies [22,33,39] included random forest, although only one [22] reported random forest as the best performing algorithm, whereas another only reported the performance of logistic regression [33], and the third study reported the average AUC across algorithms [39]. Considering significant variability across studies in top algorithms, and that only a minority of studies thus far have compared performance across two or more algorithms, model development using several algorithms may help elucidate benchmarks for certain types of input data, feature scaling methods, and specific outcomes. For instance, Wu and colleagues assessed the relative performance of their algorithm SELSER against RVM with non-SELSER-optimized features as a benchmark, showing better performance with SELSER when using clinical features alone, and EEG features [23]. Nonetheless, it is important to highlight that no singular algorithm should be expected to outperform others in all use cases. Performance is expected to largely vary based on the type of disorder, how the outcome was operationalized, the risk horizon, and the scaling of the predictors.

In terms of feature selection methods, as detailed further in Supplementary Table S3, among seven studies [15,17–19,35,39,40] incorporating peripheral blood markers as input features, only three [19,39,40] reported a method of feature selection which included non-zero β-coefficients using LASSO [19], bootstrap estimated mean decrease in Gini index within a Random Forest model [39],

and embedded feature selection (SVM, Random Forest, AdaBoost) [40]. Similarly, of the five studies [20–23,39] incorporating EEG features, four [21–23,39] reported feature selection methods which included Bonferroni adjusted significance values [21], kernel principal component analysis (PCA) [22], embedded feature selection using a sparse constraint on the number of spatial filters (22), and the highest C index scores [39].

Furthermore, among the nine studies using fMRI input features [24–32], five used a correlation based method of feature selection [14,25,27,29,30], including false discovery rate (FDR) corrected two-tailed tests with a $p<0.05$ [25] and $p<0.1$ threshold [30], a family-wise error corrected significance threshold set at $p<0.05$ (two-tailed) [14], and Pearson correlation with a significance threshold set at $p \leq 0.10$ [27], and $p < 0.05$ [29], respectively. Other feature selection methods included embedded feature selection within RVM [26], Support Vector Regression [32], and beta weights from each event during Targets, Commissions, and Rejections [34]. Similarly, among the seventeen studies (73.9%) reporting feature extraction methods, little overlap across studies was observed across studies using the same modality of input data, as further detailed in Supplementary Table S3. As such, greater continuity within feature selection and extraction methods and a comparison of multiple approaches is required in prospective studies to assess their relative efficacy to derive meaningful features of treatment response.

Quality Assessment

Overall, moderate sample sizes were used to develop predictive models of treatment response within clinical trials in psychiatry, with a median sample of 251 (n=3) in peripheral blood marker studies, 53 patients (n=5) across EEG studies, 92 patients (n=9) across neuroimaging studies, and 240.5 patients (n=10) in multimodal studies, respectively. Quality metrics were evaluated using a quality assessment instrument specific to machine learning, as described in Supplementary Table

S2. Quality scores ranged from 4/9 (44.4%) to 9/9 (100%), with the highest score in a cross-trial replication study comprising data from two large-scale trials of escitalopram and citalopram [36]. Furthermore, sixteen studies (61.5%) [17–19,21–23,25–32,37] assessed treatment outcomes in a double-blinded manner, and nineteen studies (76%) [14,15,17,19,21,23–26,28,29,31,32,34–38,40] reported all expected performance metrics including coefficient of determination and significance value within regression models, and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and AUC within classification models.

However, only six studies (23.0%) [19,21,24,28,36,40]used separate training and testing sets, with the vast majority (76.9%) instead relying on internal cross-validation to assess model performance. While several such studies lacking training and testing sets involved small sample sizes, in 52.6% of cases [15,23,25–27,29,30,35,37,38] performance was assessed using cross-validation alone although the study sample surpassed 100 patients. Similarly, of the six studies (26%) that used separate training and testing sets, only five (20%) [19,21,24,28,40]reported either the standard deviation of model accuracy or 95% confidence intervals. Furthermore, only four of twenty studies incorporating classification models [14,27,32,36] (20%) described how class imbalance was handled, where applicable, while this consideration was unclear in sixteen studies (80%) [15,17,19–22,24,28,30,31,33–35,37,39,40].

METHODOLOGICAL RECOMMENDATIONS

This is the first systematic review comprising predictive models of treatment response in randomized clinical trials in psychiatry. Throughout our review, we have identified recurrent data-driven biomarkers, algorithms, and pre-processing strategies used in predicting treatment response at an individual level. However, while several studies identified in the present review have used machine-learning models to predict treatment response using data derived from

existing large-scale randomized clinical trials, machine-learning guided interventional trials are lacking in psychiatry. Therefore, we propose a methodological pipeline to conduct prospective machine-learning guided trials according to best practices and provide strategies to improve the interpretability and generalizability of predictive models. Further discussion related to model interpretability, dealing with class imbalance, predicting adverse drug reactions in clinical trials, and calculating heterogeneity scores in patients can be observed in the supplementary material.

Machine-learning precision trials

All included studies developed models using previously collected data, which necessitates a caution of their clinical implementation without adequate prospective validation. While RCTs have provided important insights into group-level statistics, they fail to yield individualized findings or account for patient heterogeneity. As such, we advocate for a new trial design to occur following the successful completion of an RCT. We refer to this as a *machine-learning precision trial*. Using standard RCT data within machine learning models garner two major limitations: (1) The sample included in the RCTs are not fully representative of the real clinical population with a specific disorder and (2) a considerable amount of the sample size is dedicated to a placebo condition, which may be better allocated towards an active arm from a modelling perspective.

Machine-learning precision trials must therefore possess three distinct components from traditional RCTs: (1) The vast majority of participants (≤90%) are allocated to the active treatment, and a small subset of patients (≥10%) are allocated to a placebo or sham control. This allows for testing the specificity of biomarkers identified within the treatment arm; (2) greater flexibility in inclusion and exclusion criteria to increase the external validity of the trial, and reflect heterogeneous patients seen in the clinic, and (3) randomizing patients to medication

dosages in the therapeutic range known to be effective, so that machine learning models can be trained to determine more individualized dosages based on patient characteristics.

With respect to the second consideration, it is important to note that while patient idiosyncrasies are commonly observed in real-world clinical settings, such as comorbidities, are common exclusion criteria in RCTs, greater flexibility in exclusion criteria may help to provide a more realistic appraisal of the generalizability and clinical utility of machine-learning precision trials.

Furthermore, although decreasing the sample size of individuals allocated to placebo conditions is required to maximize the sample in the active arm, it may be useful to retain a small proportion of the sample (approximately 10-20%), to be given an inert substance or sham condition, to determine the specificity of features relative to placebo. Additionally, other methods can be useful to control for placebo related features, such as utilising PCA to identify the components explaining the majority (≥90%) of variance in predicting response to placebo and using a method such as multivariate adaptive regression splines (MARS) [61], where placebo related variance is imputed in the forward pass and removed from the set of candidate features in the backwards pass.

Machine-learning trials of treatment selection

Importantly, while machine learning precision trials may initially develop clinical calculators of response to a single medication, to facilitate true precision medicine in psychiatry, a focus on individual differences in the comparative effectiveness of multiple medications is required. This involves comparing multiple treatments within the same group of patients in a randomized cross-over trial, to determine the optimal medication for each individual patient to receive. In such cases, randomizing patients to different doses within a therapeutic window may not be realistic,

as very large samples would be required to generate sufficient training data of all medications within the trial, and medication dosages. To our knowledge, no such trials predicting the comparative effects across multiple treatment arms have yet been done.

Figure 1: Schematic for prospective machine learning-guided trials.

(1) A broader protocol design is used to more accurately represent heterogeneous patients seen within the clinic.

(2) A) Ninety percent (90%) of patients within the trial are assigned to an active treatment and receive a randomly selected dose within the established therapeutic range of the medication. A truncated placebo arm (10%) changes dosage of the inert substance proportional to the active treatment. This condition is used to test the specificity of data-driven biomarkers (top features).

(3) The trial continues treatment according to the duration established within phase III clinical trials. (4) Patient outcomes according to medication dosage are recorded, and common side effects are predicted at an individual level.

(5) Individualized predictive models are created and used to develop clinical calculators. The sample size of the model should be sufficient to separate into training and testing sets of adequate sizes. The exact training/testing split may vary based on sample size, however, a common threshold used within studies is allocating 70% of the sample to training, and 30% to testing, respectively. Furthermore, the size of the test set should be sufficiently large for accuracy and other metrics to be estimated with high reliability.

(6) Methods such as SHapley Additive exPlanations (SHAP) [62] are used to explain the output of predictive models and examine the effects of individual variables on model output.

(7) Optionally, randomized cross-over trials of treatment selection are conducted, where patients are assigned to one of several medications at a dosing regimen used in prior phase III trials, to predict the optimal treatment, among a candidate set at an individual level (treatment selection prediction).

Perspectives

Feature Screening and Extraction

Several studies [19,25–27,29–32,34] included in the present review utilized high-dimensional features, defined as a significantly greater number of predictors (p) relative to the number of patients (n) in the training sample. In certain cases, such as when dealing with genome-wide genetic data, the number of candidate features can grow exponentially with the sample size, resulting in ultra-high-dimensional data [63]. In these cases, feature screening procedures, involving rank ordering features and significantly reducing dimensionality, can be useful prior to standard feature selection methods. While there are several available methods of feature screening, as explained elsewhere [64], model-free approaches are particularly useful in machine learning models with many candidate features, and limited evidence suggesting a parametric distribution of features.

For instance, Zhu and colleagues [65] developed a model-free feature screening method for ultra-high-dimensional data that is computationally efficient and robust to outliers, which utilises hard and soft thresholding strategies to obtain a cut-off point that separates active (relevant) and inactive (redundant) predictors for a given outcome [65]. More recently, Li and colleagues (64) developed a feature screening method for ultra-high-dimensional data where an outcome (e.g., post-treatment symptom severity) is missing at random. This approach is based on an adjusted Spearman rank correlation, in conjunction with a nonparametric imputation technique. Of note,

this method is developed on the assumption that the candidate predictors are continuous, and the input features are not grouped data [66]. Additionally, another recent method by Guo and colleagues [67] abandons hard rules, in favour of a method incorporating data-adaptive threshold selection, which can control the per family error rate and false discovery rate under certain conditions, while retaining all important features.

Apart from feature screening measures, there are several newly developed feature engineering approaches that may be useful in predictive models of treatment response and selection, defined broadly as developing features from raw data or creating new variables from original variables. While an exhaustive overview of feature extraction methods is outside the scope of this review, a survey of feature extraction techniques for machine learning models can be found elsewhere [68].

In terms of recently developed methods for high dimensional data, Bonidia and colleagues [69] developed a method of feature extraction for biological sequencing data based on mathematical features, including six numerical mapping techniques with Fourier transform, Tsallis and Shannon entropy, and graphs (complex network). This mathematical method was compared against biological feature extraction methods (e.g., LncRNA-ID, IncRScan-SVM), and models using mathematical feature extraction methods reported the best performance (89.01-96.06% accuracy) across RNA classification tasks with an improvement of 3.28% and 3.01% across tasks relative to biological feature extraction methods alone. Of note, a hybrid model combining both mathematical and biological approaches of feature extraction improved accuracy, suggesting that merging features may improve the predictive performance of classification tasks. This method of merging biological and mathematical feature extraction methods may be particularly useful for prospective trials of treatment response and treatment selection using genome-wide SNPs and whole-blood RNA as input features [69].

Furthermore, Barandas and colleagues [70] developed Time Series Feature Extraction Library (TSFEL), a user-friendly Python package that provides a comprehensive list of feature extraction methods for time series data, across temporal, statistical, and spectral domains. Of note, TSFEL also provides a systematic way to record inputs, the dataset, metadata, and feature extraction parameters, for reproducibility. Toolboxes such as this may be particularly useful to extract features from EEG, and time-series fMRI data in prospective models [70].

Algorithms, hyperparameter tuning, and stacked generalization

Furthermore, an important consideration in model development is comparing multiple algorithms (e.g., linear, tree-based, and kernel methods) to assess their relative performance in predicting an outcome of interest. Approximately 35% of studies (n=8) included in the review compared model performance using at least two different algorithms. Apart from the standard comparison across algorithms, stacked generalization provides an alternative ensemble method to combine the predictions of two or more machine learning algorithms, while using another algorithm to learn how to combine their outputs [71]. As described elsewhere [72], stacking can improve model performance over any single model contained in the ensemble. Additionally, stacking differs from the traditional bagging and boosting ensemble methods in that it typically uses different models that combine predictions from contributing models, rather than a series of decision trees, or models that comprise weak learners building upon the prediction of previous models, respectively. While no studies included in the present review utilized stacked generalization in model development, this approach may be useful to improve model accuracy in prospective studies.

Moreover, hyperparameter tuning, which involves selecting the optimal set of hyperparameters for a given model, remains an important consideration in model development. While many software packages have default hyperparameter settings during cross-validation, searching the hyper-parameter space for the lowest loss-function, or best cross-validation score, is recommended. Although an exhaustive search of the hyperparameter space is often computationally infeasible, there are several available methods such as a manual grid search, collaborative hyperparameter tuning, and Bayesian optimization.

Importance of precision in performance estimates

In the context of classification models, it is important to highlight that uncertainty estimates should be considered when evaluating model accuracy and other common performance metrics such as sensitivity and specificity. For instance, while a specific model may show a reasonable accuracy, if a large range is observed between the upper and lower bounds of the 95% confidence interval, it is plausible that the model may be too imprecise to reasonably predict treatment response or selection in a prospective trial. Therefore, in the absence of uncertainty estimates such as confidence intervals, it is imperative that model performance is interpreted with necessary caution. It is also worth noting the inherent difficulty in estimating the variability of cross-validated performance metrics [73]. Additionally, many other fields successfully use cross-validation as a basis for choosing between different models or tuning regularization parameters for a model, rather than taking its performance estimate at face value [74].

Within the current review, only 6 of 26 studies (23.0%) [19,21,24,28,36,40]incorporated training and testing sets during model development, allowing for a comparison of uncertainty estimates across these models. Among them, only five studies (19.2%)[19,21,24,28,40] reported either the standard deviation of model accuracy or 95% confidence intervals. Further information can be

found in Supplementary Table S3. As such, there remains an urgent need for prospective models to report the uncertainty estimates of performance metrics.

Performance metrics and their implications within precision medicine

Apart from the important considerations of uncertainty estimates, there is a need to consider the relationship between performance metrics and their implications within precision medicine. Common methods of evaluating the performance of ML classification models across studies contained within this review include accuracy, sensitivity, specificity, PPV, NPV, and AUC.

Although these metrics all provide useful information to evaluate the potential utility of the model, it is important to consider the relationship between them and their likely expected benefits for treatment selection. For instance, seventeen of twenty-six studies (65.38%) [14–17,19–22,27,28,30,31,35,36,39,40,75] used a binary classification task to predict clinical response vs. non-response to a specific intervention. In this instance, the sensitivity of the model corresponds to its ability to correctly identify patients who will respond to the intervention (true positive), while specificity relates to the ability to identify patients who are likely to be non-responders (true negative). Additionally, PPV and NPV provide insight into the prevalence of the outcome, and indicate the likelihood of clinical response, or non-response, in the case of a positive or negative result, respectively.

Although the ideal threshold between sensitivity and specificity largely depends on the baseline rates of treatment efficacy for a given intervention, it is important to highlight that reasonable balanced accuracy does not necessarily translate into a model with clinical utility or scalability. For example, a binary classification model with a balanced accuracy of 67.5% in predicting response vs nonresponse to clozapine, corresponding to 45% sensitivity (true positive) and 85%

specificity (true negative), shows worse performance than random chance at identifying whether a given patient will meet a pre-specified threshold for clinical response to the medication. While clozapine has been shown to be an effective treatment in psychotic disorders [76], it also facilitates a host of undesirable side effects, including drowsiness, hypersalivation, and constipation [77]. As such, this hypothetical model will perform extremely poorly in identifying which patients will respond to clozapine, and the associated predictors lack discriminative capabilities in this regard. In other words, important features, or biomarkers, within this model provide a signal for identifying whether a patient will not respond to clozapine but fail to provide meaningful signals for therapeutic response.

Conversely, even with an 85% specificity (true negative), this model will misclassify patients as non-responders in 15% of cases. This misclassification error, or number of false negatives, scales proportionally to the overall sample size, leading to many individuals prescribed a medication with many adverse side effects that will ultimately be ineffective when implemented clinically.

Therefore, when evaluating performance thresholds to ascertain whether a given model is sufficiently accurate to make a useful impact in selecting treatments, it is important to consider the expected efficacy of the intervention, the therapeutic safety profile, and whether the proportion of true positives and true negatives within a model provide a meaningful performance threshold for a given disorder and intervention. Moreover, metrics such as PPV and NPV provide useful context into the prevalence of a given outcome, and should be considered alongside sensitivity, specificity, and AUC.

Novel features in prospective models of treatment response and selection

Throughout the review, models of treatment response within randomized clinical trials have been developed using peripheral blood markers comprising SNPs and fatty acid composition, resting-state EEG, resting-state, and task-specific fMRI, as well as multimodal data comprising combinations of clinical, genetic, EEG, and fMRI features. Besides the approaches used in the literature thus far, there are several types of features that may be useful to incorporate in prospective models of treatment response and selection.

In terms of whole-blood peripheral biomarkers, next-generation sequencing methods such as RNA sequencing (RNA-seq) can be used to identify gene expression markers that are predictive of treatment response. For instance, Nøhr and colleagues [78] used data from a placebo-controlled trial comprising 184 patients treated with either vortioxetine or placebo for MDD, and using blood samples collected with PAX gene tubes, identified three novel genes whose RNA expression levels at baseline and week 8 were significantly (FDR <0.05) associated with treatment response after 8 weeks of treatment. However, they did not identify any genes that were differentially expressed between placebo and vortioxetine groups [78]. More recently, new low-cost, portable high-throughput single-cell RNA sequencing methods have been developed, which have been used for cell-specific biomarker discovery [79]. Importantly, new feature selection methods are available for biomarker discovery using sparse single cell data. For example, it was shown that a probabilistic generative model can reduce the high-dimensional space in single-cell gene expression data and provide uncertainty estimates [80].

With respect to neurophysiological measures such as EEG, new multimodal techniques have been developed, such as combining TMS with EEG, to directly and non-invasively explore cortical reactivity with improved temporal resolution [81]. This allows for examining several types of features, including cortical excitability, cortical inhibition, cortical oscillations, and the

balance between excitation and inhibition within the cortex in response to TMS pulses. This technique may be particularly useful in randomized trials of rTMS, by measuring baseline brain neurophysiology and mid-treatment. For instance, in a study by Voineskos and colleagues [82], N45 amplitude measured using TMS-EEG over the DLPFC was shown to discriminate individuals with depression from healthy controls with 76.6% accuracy (80% sensitivity, 73.3% specificity, AUC: 0.829) [82].

In terms of functional neuroimaging, functional near-infrared spectroscopy (fNIRS) is a method that uses near-infrared light to estimate cortical hemodynamic activity in response to neural activity [83]. While fNIRS has several remaining limitations [84], such as a depth sensitivity of approximately 1.5 cm, and a spatial resolution up to 1 cm, it has recently been used to dichotomize patients with MDD from healthy controls, with frontal region integral values correctly classifying 75.2% of patients with MDD, and 74.3% of healthy controls, respectively [85]. However, it remains to be investigated whether this has utility in identifying predictors of treatment response between individuals within the same diagnostic category.

Furthermore, in terms of low-cost features that may be predictive of treatment response, there is increasing interest in the use of speech-based biomarkers adopted using smartphone technology [86]. For instance, in a study by Mundt and colleagues [87] comprising 105 adults with MDD, it was found that baseline and week 4 speech markers could predict responder vs non-responder status to sertraline at week 4 with a sensitivity estimate of 70.6% and specificity estimate of 79.2%, respectively. Moreover, six vocal acoustic measures were found to significantly correlate with depressive severity scores, as measured using the Quick Inventory of Depressive Symptomatology - Clinician Rating (QIDS-C) scale. This included total pause time, pause variability, percent pause time, speech/pause ratio, and speaking rate [87].

CONCLUSION

While RCTs and evidence-based medicine have facilitated undeniable advancements in patient care, personalised interventions remain a critical need in mental health [88]. Machine-learning precision trials may help us move away from the "one size fits all" assumption of current trials by including patient heterogeneity in individualized models. Similarly, assigning patients to a randomly selected dose in the established therapeutic range, while keeping important considerations such as body weight and contraindications in mind, may facilitate useful algorithms to titrate medications with greater granularity. However, this will require large sample sizes, and appropriate training, testing, and external validation prior to clinical implementation.

Importantly, although treatment response prediction has utility in prognosticating whether a patient will respond to a specific intervention, they cannot determine the optimal treatment option for a specific patient. As such, machine-learning guided models of treatment selection, evaluating individual differences in comparative effectiveness across the same group of patients, are required to facilitate precision psychiatry.

Conflict of interest statements

Devon Watts reports a PhD fellowship from the Canadian Institute of Health Research (CIHR), outside the submitted work. Diego Librenza-Garcia, Pedro Ballester, Bruno Jaskulski Kotzian, Jessica Yang, Benício Frey, Luciano Minuzzi, Benson Mwangi, and Ives Cavalcante Passos report no biomedical financial interests or potential conflicts of interest. Flávio Kapczinski has received grants/research support from AstraZeneca, Eli Lilly, Janssen-Cilag, Servier, NARSAD, and the Stanley Medical Research Institute; has been a member of the speakers' boards of AstraZeneca, Eli Lilly, Janssen and Servier; and has served as a consultant for Servier.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful and constructive comments on a prior version of this manuscript. There is no financial support or grant sources to disclose in association with this work.

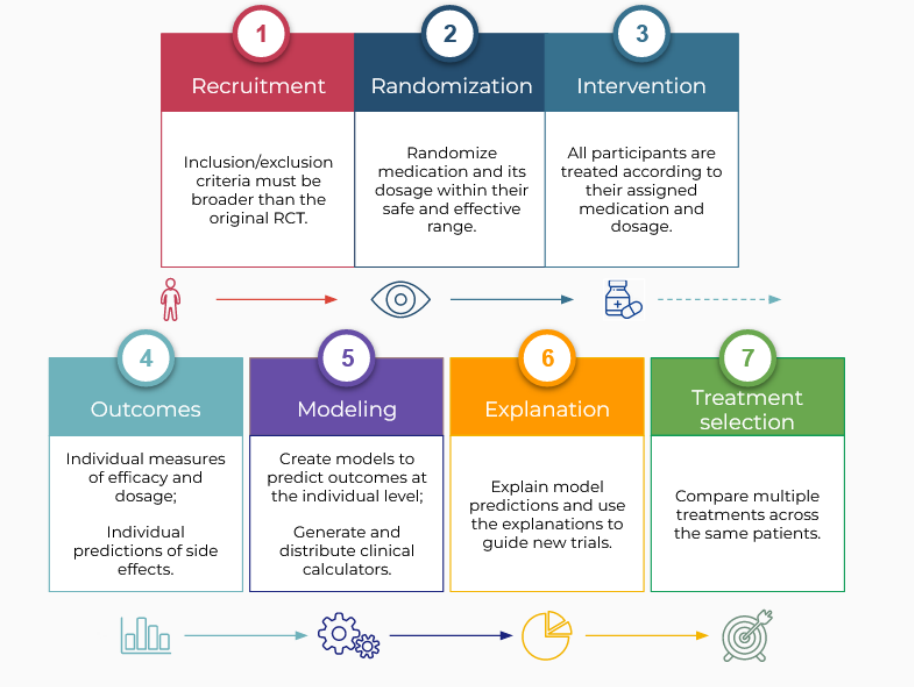


Figure 1: Schematic for prospective machine learning-guided trials.

(1) A broader protocol design is used to more accurately represent heterogeneous patients seen within the clinic.

(2) A) Ninety percent (90%) of patients within the trial are assigned to an active treatment and receive a randomly selected dose within the established therapeutic range of the medication. A truncated placebo arm (10%) changes dosage of the inert substance proportional to the active treatment. This condition is used to test the specificity of data-driven biomarkers (top features).

(3) The trial continues treatment according to the duration established within phase III clinical trials.

(4) Patient outcomes according to medication dosage are recorded, and common side effects are predicted at an individual level.

(5) Individualized predictive models are created and used to develop clinical calculators. The sample size of the model should be sufficient to separate into training and testing sets of adequate sizes. The exact training/testing split may vary based on sample size, however, a common threshold used within studies is allocating 70% of the sample to training, and 30% to testing, respectively. Furthermore, the size of the test set should be sufficiently large for accuracy and other metrics to be estimated with high reliability.

(6) Methods such as SHapley Additive exPlanations (SHAP) [61] are used to explain the output of predictive models and examine the effects of individual variables on model output.

(7) Optionally, randomized cross-over trials of treatment selection are conducted, where patients are assigned to one of several medications at a dosing regimen used in prior phase III trials, to predict the optimal treatment, among a candidate set at an individual level (treatment selection prediction).

First author, year	Sample size and diagnosis[1,2]	Clinical Trial	Outcome	Machine learning model	Data utilized	Top Data-Driven Biomarkers (Features)	Accuracy (95% CI)	Additional Performance Metrics
STUDIES USING PERIPHERAL BLOOD MARKERS								
Amminger, 2015	81 individuals at ultra-high risk of psychosis -27 males -54 females	ω-3 PUFAs vs. placebo	Responder vs non-responder ≥ 15-point increase in GAF score classified as responders.	GPC	Erythrocyte fatty acid composition of the phosphatidylethanolamine quantified via capillary gas chromatography. ALA, EPA, DPA, DHA, LA, AA, and NA were examined	ω-3 response: 1. Nervonic acid 2. margaric acid 3. arachidonic acid Placebo response: 1. Erucic Acid 2. Arachidonic acid 3. Docosahexaenoic acid	ω-3: 86.7% placebo: 79.6%	ω-3: sensitivity - 86.7%; specificity: 86.7% placebo: sensitivity 83.3%; specificity: 75%
Hou, 2015	251 patients with AUD	11 weeks of Ondansetron (5-HT3 receptor antagonist) vs placebo	Percentage of heavy drinking days (PHDD)	IT VT LR	Genotyping of long and short alleles of the functional insertion-deletion polymorphism (5'-HTTLPR) in the promoter region of the SLC6A4 gene. A total of 21 genetic polymorphisms were	VT model: 1. rs1150226-GG 2. rs1176719-AG 3. PHDD_base <0.883 IT model: 1. rs1150226-AG 2. rs1176719-AG 3. Onset age ≥ 23	N/A	Mean difference of PHDD - IT subgroup: 17.2% - VT subgroup: 21.8% (Table 1) subgroup comparison for % of patients with ≤1

					considered as predictors.			heavy drinking day in month 3 of follow-up -TRM (n=57) OR 5.0, p=0.015 - VT (n=88) OR 3.8, p=0.017 - IT (n=118) OR 2.1, p=0.05
Maciukiewicz, 2018	450 patients with MDD	Enrolled in one of three clinical trials where each patient received duloxetine or placebo for up to 8 weeks	Responders vs Non-responders ≥ 50% reduction in MADRS total score at endpoint Remission vs non-remission ≤10 MADRS total score at	CART SVM	571,054 SNPs generated using an Infinium PsychArray-24 Kit	19 SNPs rs2036270 rs7037011 rs1138545 rs1107372 rs11136977 rs11581838 rs11843926 rs1347866 rs16932062 rs19999223	Response CART: 55-57% SVM: 64-66% Remission CART: 45-51% SVM: 51-52%	Response Sensitivity (CART): 71-75% Specificity (CART): 15-17% Sensitivity (SVM): 87-89% Specificity (SVM):

			endpoint			rs2710664 rs2710664 rs39185 rs4520243 rs46858655 rs4777522 rs4954764 rs60230255 rs6550948 rs972016		7-9% Remission Sensitivity (CART): 45-51% Specificity (CART): 33-51% Sensitivity (SVM): 58-59% Specificity (SVM): 41-46%
First author, year	**Sample size and diagnosis[1,2]**	**Clinical Trial**	**Outcome**	**Machine learning model**	**Data utilized**	**Top Features**	**Accuracy**	**Additional Performance Metrics**

STUDIES USING ELECTROENCEPHALOGRAPHIC MEASURES								
Al-Kaysi, 2017	10 patients with MDD	15 sessions of tDCS or sham over 3 weeks, followed by an optional 3-weeks of open-label tDCS	Responder vs non-responder ≥ 50% reduction in MADRS scores from baseline to treatment session 15 or 23 (Assessment at week 23 was part of the open-label trial.	SVM ELM LDA	Continuous eyes-closed resting-state EEG over 10 minutes. Average PSD in conventional EEG frequency bands (delta, theta, alpha, beta, gamma). Alpha asymmetry in the frontal, central, and parietal cortices.	<u>Frontal</u> AF2-AF8 - 71% AF8-F9 - 71% AF-AF8 - 70% <u>Central/Parietal</u> T8-C1 - 73% T8-CpZ - 71% T8-Cz - 70% <u>Parietal/Occipital</u> Pz-P2 - 68% Pz-PO4 - 62% Pz-PO3 - 61% <u>All regions</u> FC4-AF8 - 76% T8-C1 - 73%	76% Best performance using FC4-AF8 electrode pairs (76%) Channel Tp9 performed best for predicting responder vs nonresponder status (71±11%	NA

						T8-CPz - 71%		
Cao, 2019	55 patients with treatment-resistant depression	Double-blind placebo-controlled trial (1:1:1): 1) 0.5 mg/kg ketamine 2) 0.2mg/kg ketamine 3) Normal saline	Responders vs non-responders ≥ 45% reduction in HDRS-17 from baseline to 240 min post-treatment	LDA NMSC kNN PARZEN PERLC DRBMC SVM *Radial kernel*	*Resting-state EEG functional connectivity measures* EEG Power EEG Alpha Asymmetry	Responders in the 0.5mg/kg ketamine group showed lower relative EEG theta and lower alpha power ($p < 0.05$) Responders in the 0.2mg/kg showed significantly weaker relative EEG power in the theta band on the Fp2 channel than non-responders	78.4% *Best performance using SVM with a radial kernel*	Sensitivity: 79.3% Specificity: 84.2% Recall: 78.5% Precision: 87.0% F1 score: 52.6%
de la Salle, 2020	47 patients with MDD	12-week double-blind trial of: 1) escitalopram 2) bupropion 3) escitalopram + bupropion	Responders vs. Non-responders (≥50% reduction in MADRS scores from baseline to posttreatment) Remitters/Non-remitters ≤10 MADRS at	LR	*EEG Theta cordance* *EEG middle right frontal*	Best performance using change in prefrontal theta cordance	Response 74-81% *Best performance using change in PF cordance* Remission 51-70%	AUC: 0.85 Sensitivity: 70% Specificity: 85% PPV: 0.95 NPV: 0.76 *Remission* (ΔPF): AUC: 0.66 Sensitivity: 65% Specificity: 74% PPV: 65% NPV:

			posttreatment					74% *Response* (ΔMRF): AUC: 0.80 Sensitivity: 70% Specificity: 95% PPV: 95% NPV: 76% *Remission* (ΔMRF): AUC: 0.59 Sensitivity: 93% Specificity: 31% PPV: 39% NPV: 91%
Jaworska, 2019	51 patients with MDD	12-week double-blind trial of: 1) escitalopram 2) bupropion 3) escitalopram + bupropion	Responders vs Non-responders ≥50% MADRS score reduction from baseline	RF AdaBoost SVM CART MLP Gaussian naive Bayes *Best overall performance*	*Pre-treatment rs-EEG* eLORETA EEG Band Power	Alpha 2 - eLORETA Alpha 2 - EEG band power Alpha 1 - eLORETA Theta - EEG band power Delta - eLORETA Beta - EEG band power	NA	Alpha 2 - eLORETA AUC: 0.585-0.803 Alpha 2 - EEG band power AUC: 0.689-0.783 Alpha 1 -

				using RF		*The most predictive features were EEG delta power at week 1 at T8 followed by power at CP6.*		eLORETA AUC: 0.635-0.756 Theta - EEG band power AUC: 0.664-0.752 Delta - eLORETA AUC: 0.569-0.718 Beta - EEG band power AUC: 0.689-0.783
Wu, 2020	309 patients with MDD EMBARC study (n=228) tested in two independent samples (n=72) and (n=24)	8-week double-blind trial of sertraline or placebo	Pre- minus post-treatment difference in HAMD17 scores	SELSER *Algorithm developed in the current study* RVM used as a	Pre-treatment resting-state EEG (eyes open/eyes closed) SELSER - neural signals drawn from θ and α frequency bands	For the sertraline arm, only signals from the resting-eyes open condition (alpha band) were significantly predictive of treatment score change during cross-validation.	R^2 0.60 *Sertraline* R^2 0.41 *Placebo* *Different*	Less rsEEG-predicted HAMD17 change with sertraline was associated with greater response to 1-Hz rTMS on the DASS (rsEEG-predicted HAMD17 sertraline change x

	respectively			comparator	SELSER optimizes a sparse set of spatial filters that map EEG signals to a latent space, and then relates the band powers of the latent signals to the treatment outcome via a linear regression model.		*important features in sertraline and placebo models*	time interaction: F (1,128) =9.02, P=4 X 10-3)
First author, year	**Sample size and diagnosis[1,2]**	**Clinical Trial**	**Outcome**	**Machine learning model**	**Data utilized**	**Top Features**	**Accuracy**	**Additional Performance Metrics**
STUDIES USING NEUROIMAGING								
Braund, 2022	226 patients with MDD	Patients randomized in a 1:1:1: ratio to escitalopram, sertraline, or extended-release venlafaxine for 8 weeks (iSPOT-D)	Responders vs Non-responders >50% reduction in $HDRS_{17}$ or $QIDS\text{-}SR_{16}$. Remission vs non-remission	SVM	Baseline intrinsic functional connectivity between each pair of 436 brain regions. 19 connections across 30 brain regions that were associated with neuroticism (total	Most important edges consisted of connections between the somatomotor and limbic networks, limbic and executive control networks, executive control, and dorsal attention networks, and somatomotor and visual networks.	75% (95% CI: 57.8-87.9%)	Sensitivity: 62.5% Specificity: 85.0% PPV: 76.9% NPV: 73.9%

			$HDRS_{17}$ score <7 or QIDS-SR_{16} score <5 at week 8		NEO-FFI scores) were used in final model			
Fan, 2020	200 unmedicated patients with MDD	EMBARC trial All patients randomly assigned to 8 weeks of either sertraline or placebo (up to 200 mg daily)	Percentage change in HAMD-17 scores before and after 8 weeks of treatment	CPM	*rs-fMRI* functional connectome fingerprints	Enhanced treatment response was predicted by lower pretreatment connectivity between the executive and sensorimotor and salience modules, but increased connectivity between the DM modules and the rest of the brain.	Sertraline or placebo: r=0.19 *No difference in features between sertraline and placebo models*	Secondary analyses also identified pretreatment CFP at higher resolution (A424), which significantly predicted percentage of symptom improvement (r=0.19, CV=10, iterations=1000, p=0.02), while pretreatment CFP at lower resolution (AA-24) showed a trend on the prediction (r=0.14, CV=10, iterations=1000, p=0.08)

Fonzo, 2019	251 unmedicated patients with MDD	EMBARC trial All patients randomly assigned to 8 weeks of either sertraline or placebo (up to 200 mg daily)	Pre-minus-post change in HAMD-17 scores	RVM	fMRI during an emotional conflict task	Important features specific to the sertraline RVM model included the right insular lobe and right middle temporal gyrus. Features that predicted treatment outcome across study arms included the left anterior cingulate cortex/superior medial gyrus, and both hemispheres of the anterior cingulate cortex.	Sertraline: r=-0.49, P<0.001 Placebo: r=-0.06, P=0.48	Interesting, an RVM model trained on emotional conflict regulation brain activation data in the placebo arm to predict placebo outcome did not yield significant correlations between model-predicted symptom changes and observed symptom changes in either the placebo or sertraline arms (r=0.11, P>0.20)
Klöbl, 2020	35 patients with MDD	Randomized, double-blind, cross-over trial of 8 mg intravenous citalopram or placebo	Responder vs non-responder ≥50% reduction in HAM-D scores from baseline to posttreatment	Linear regression with robust "bisquare" weighting	*fMRI* Network-based statistical analysis	Top predictors included voxels in the ventral attention (VA; e.g., anterior midcingulate cortex, left superior temporal and supramarginal gyrus, insula, eye fields), default mode (DM; e.g., frontal cortex, anterior and	NA	Response AUC = 0.68 Remission AUC = 0.73

			(median of 10 weeks) Remission vs non-remission HAMD ≤ 7			posterior cingulate cortex, precuneus) and fronto-parietal (FP; e.g., frontal and prefrontal cortex, anterior cingulate cortex) networks.		
Koutsouleris 2017	92 patients with SCZ	Randomized to either active (N=45) or sham (N=47) 10-Hz rTMS applied to the left DLPFC 5 days per week for 21 days	Responders vs non-responders ≥ 20% improvement in PANSS-NS between baseline and day 21	SVM	*sMRI* Total intracranial, gray matter, white matter, and cerebrospinal fluid volume	Largest feature weights (overall mean standard error) included global CSF volume and total intracranial volume	Active rTMS: 61.5-71.4%	Sensitivity: 61.5-70.8% Specificity: 66.7-71.1% PPV: 54.5-75.0% NPV: 54.5-75.0%
Nemati, 2020	258 patients with MDD EMBARC trial (n=202) Independent	8-weeks of daily oral placebo or sertraline	Percentage change in HAMD-17 scores before and after 8 weeks of treatment	CPM	*rs-fMRI* (Connectome fingerprints)	Whole brain NRS-PM predicted antidepressant response across AA-4 to AA-150 architectures (following FDR correction), with a peak at AA-58 (r=0.27, CV=10, iterations=1000, p=0.003)	r= 0.25-0.29	N/A

	RCT of ketamine, placebo, or (n=56)					Positive predictive edges peaked at AA-58 (r=0.29, CV=10, iterations=1000, p=0.001) and the negative predictive edges peaked at AA-26 (r=0.25, CV=10, iterations=1000, p=0.003)		
Nord, 2019	39 unmedicated patients with MDD	Double-blind trial of 8-weeks of real (N=20) or sham (N=19) tDCS Immediately following each tDCS session, patients received a 1-h CBT intervention for depression n-back working memory task	Responder vs non-responders ≥50% reduction in HAM-D scores from baseline to posttreatment Remission vs non-remission HAMD ≤ 7	LDA	*fMRI* whole-brain flexible factorial analysis ROIs for the emotional processing task included the left and right amygdala, subgenual anterior cingulate cortex, and L-DLPFC	Baseline L-DLPFC activation was shown to discriminate responders from non-responders with an AUC of 0.856. Of note, this same pattern of activation did not discriminate responders from non-responders in the sham condition (AUC = 0.417).	N/A	Response to active tDCS AUC: 0.856 Response to sham tDCS AUC: 0.417

		performed during stimulation						
Sarpal, 2016	41 patients with first-episode schizophrenia	Double-blind randomized controlled treatment with either risperidone or aripiprazole for 52 weeks (18 patients with first-episode schizophrenia treated with aripiprazole 22 treated with risperidone)	Responder vs non-responder Responders defined as two consecutive visits with a CGI improvement score ≥1 and a rating ≤3 on the following BPRS items: conceptual disorganization, grandiosity, hallucinatory behavior, and unusual thought content	Cox-regression	*rs-fMRI* Functional Connectivity Analyses Voxel-Wise Survival analysis Striatal Connectivity Index	The insular cortex, opercular cortex, anterior cingulate, thalamus, orbitofrontal cortex, and posterior cingulate were regions that frequently appeared on the list of predictive connections with the striatum. In posterior regions, greater connectivity with striatal subdivisions at baseline were associated with better subsequent treatment response. In more frontal regions, by contrast, lower striatal connectivity of these nodes at baseline was	77.5%	Sensitivity: 80% Specificity: 75% PPV: 76% NPV: 79%

						associated with better subsequent response.		
Yip, 2019	74 patients with cocaine-use disorder	Randomized controlled trial of behavioral therapy plus galantamine or placebo	Abstinence during treatment was determined using biweekly urine testing and defined as the percentage of urine negative for cocaine provided during treatment. A classification model was also used, dichotomizing patients by the presence or absence of any cocaine-negative result	CPM	fMRI during a monetary delay task	Highest-degree nodes (i.e. nodes with the most connections) for the positive network included a prefrontal node with connections to limbic, temporal, parietal, cerebellar, and other prefrontal nodes, and a temporal node with connections to limbic, parietal, motor, and prefrontal nodes. - Highest-degree nodes for the negative network also included a temporal node with connections to limbic, parietal, and prefrontal nodes as well as with connections to cerebellar and subcortical nodes.	r=0.36 Accuracy: 64%	Sensitivity: 35% Specificity: 82%

First author, year	Sample size and diagnosis[1,2]	Clinical Trial	Outcome	Machine learning model	Data utilized	Top Features	Accuracy	Additional Performance Metrics
STUDIES USING MULTIMODAL DATA								
Ambrosen, 2020	138 first episode SCZ	Three patient cohorts randomized to either: 1) Risperidone or zuclopenthixol for 3 months 2) Quetiapine for 6 months 3) Amisulpride for 6 weeks	Short term treatment response - relative change in PANSS total score from baseline to short-term follow-up Long-term treatment response - poor response was defined as any of the following: 1) clozapine prescription, 2)	LR Naive Bayes RF DT SVM k-NN	*Clinical, EEG and sMRI data* WAIS-III CANTAB EEG during the Copenhagen Psychophysiology Test Battery Cortical thickness, surface area, mean curvature	NA	Long-term treatment response 50.3%	Short-term treatment response NMSE = 0.96

			Eligibility for clozapine, 3) polypharmacy >90 days of treatment with at least two different antipsychotics					
Athreya, 2019	1030 white outpatients with MDD PGRN-AMPS (n=398) STAR*D (n=467) ISPC (n=165)	8-weeks citalopram or escitalopram PGRN-AMPS STAR*D ISPC	Responders vs Non-responders ≥ 50% reduction in HDRS or QIDS from baseline to post-treatment Remission vs non-remission HDRS ≤ 7 or QIDS ≤ 5	RF	Six SNPs in or near TSPAN5 (r10516436), ERICH3 (rs696692), DEFB1 (rs5743467, rs2741130, and rs2702877), and AHR (rs17137566) genes. 26 clinical and sociodemographic variables (including age, BMI, and plasma drug levels)	Men HDRS - response TSPAN5 DEFB1_1 DEFB1_2 AHR HAMD baseline ERICH3 DEFB1_3 Men HDRS - remission HAMD baseline DEFB1_2 DEFB1_1	Response 66-88% Remission 66-86%	Response Sensitivity: 0.68-0.90 Specificity: 0.63-0.85 PPV: 0.82-0.93 NPV: 0.51-0.79 AUC: 0.7-0.9 Remission Sensitivity: 0.59-0.90 Specificity: 0.71-0.84 PPV: 0.67-0.84

						AHR TSPAN5 ERICH3 DEFB1_3 Women HDRS - response DEFB1_1 HAMD baseline DEFB1_2 TSPAN5 AHR ERICH3 DEFB1_3 Women HDRS - remission HAMD baseline DEFB1_2 DEFB1_1 AHR		NPV: 0.759-0.87 AUC: 0.75-0.90

						TSPAN5 ERICH3 DEFB1_3		
Crane, 2017	49 patients with MDD medication free for at least 90 days from an SSRI or SNRI and at least 30 days from all other medications (including birth control)	Open-label treatment with escitalopram or duloxetine for 10 weeks	Percentage change in Hamilton Depression Rating Scale pre- to post-treatment	LR	Go/No-go test and fMRI ICA beta weights haemodynamic response function	Two event-related component beta weights were significant predictors of treatment response during commission errors, Components 24 and 25, and survived FDR correction. More HRF-based activation during Commission errors was observed in the right ventrolateral PFC of component 11; in the dorsal ACC of Component 24, and in four clusters of Component 25, including the rostral dorsal ACC and left medial PFC, all predicted poorer treatment response.	84%	Sensitivity: 84.2% Specificity: 80.0%

Fonzo, 2017	66 patients with PTSD	RCT of immediate treatment with prolonged exposure therapy or treatment waitlist (10 week) Sessions took place 1-2 times per week, for a total of 9-12 sessions (90 min each)	Remission vs non-remission Post-treatment CAPS ≤ 20	LDA	3-T GE Signa scanner (T1-weighted image) Emotional Reactivity Task Emotional Conflict Task Gender Conflict Task Reappraisal Task Baseline clinical features and treatment arm	Top features (*Emotional conflict*) - R Superior Frontal Gyrus/Middle Frontal Gyrus/Inferior Frontal Gyrus (Pars Triangularis) - L/R Anterior Cingulate/Middle Cingulate - R Superior Frontal Gyrus/Middle Frontal Gyrus - L Insula Lobe - R Superior Frontal Gyrus/Middle Frontal Gyrus - L Amygdala - L Superior Frontal Gyrus - R Superior Frontal Gyrus - L Inferior Temporal Gyrus/Middle Temporal Gyrus - L Anterior Cingulate/Middle	79.5%-97.7% *Best performance in combined mode (a-priori voxelwise and whole-brain exploratory analysis of conscious fear vs neutral)* *Leave-one-out classification accuracy is likely optimistic, as the authors note, "the predictive accuracy of these models is likely higher than what would be expected in an independent cohort of participants given that the same sample of*	A-priori voxelwise analysis of conscious fear vs neutral PPV=1.00 NPV=0.74 Whole-brain exploratory analysis of conscious fear vs neutral PPV=0.90 NPV=0.79 Combined model PPV=1.00 NPV=0.94

						Cingulate/Superior Medial Gyrus - L Inferior Temporal Gyrus - R Middle Frontal Gyrus/Superior Frontal Gyrus - L Inferior Frontal Gyrus (Pars Triangularis) - L Superior Frontal Gyrus - L Angular Gyrus - R Anterior Cingulate/Middle Cingulate - L Middle Frontal Gyrus - L Superior Frontal Gyrus - L Angular Gyrus/Inferior Parietal Lobule - L Inferior Temporal Gyrus - R Cerebellum - L Middle Temporal Gyrus	*participants was utilized to train the model and test its predictive accuracy."*	

						- R Cerebellum - L Middle Temporal Gyrus - R Middle Frontal Gyrus Emotional Connectivity Task - R Middle Frontal Gyrus/Superior Frontal Gyrus - R Superior Frontal Gyrus - L Middle Frontal Gyrus - R Superior Frontal Gyrus - R Anterior Cingulate/Middle Cingulate - L/R Olfactory Cortex/ Anterior Cingulate/Caudate Nucleus/Olfactory Cortex/Anterior - L/R Cingulate/Caudate Nucleus		

						Emotional vs Gender Task - LR Olfactory Cortex/Anterior Cingulate/Caudate Nucleus		
Lee, 2018	259 patients with SCZ	Clinical Antipsychotic Trials of Intervention Effectiveness - patients randomized to one of five antipsychotic medications for 18 months	Good vs Poor Outcome Good (if PANSS total decreased) vs. Poor (otherwise)	LGEM	53 clinical and sociodemographic variables, including BMI, heart rate, weight, and weight Top 25 SNPs from the GWAS for schizophrenia in the CATIE study 13 SNPs (rs10803138, rs11682175, rs6704641, rs6704768, rs215411, rs1106568, rs12522290, rs4129585,	*Feature importance not reported*	(range) Accuracy: 55-66% *(Greatest accuracy observed in Ziprasidone model)*	Specificity: 52-74% Sensitivity: 52-61% PPV: 49-86% NPV: 53-77%

					rs2514218, rs2239063, rs4702, rs12325245, and rs9636107) from the 128 genome-wide significant associations for schizophrenia identified by the Schizophrenia Working Group of the Psychiatric Genomics Consortium			
Joyce, 2021	375 outpatients with MDD PGRN-AMPS (n=264) CO-MED (n=111)	PGRN-AMPS 8-week clinical trial randomized to either: 1. escitalopram (10/mg day) 2. citalopram (20/mg day) CO-MED 7-month clinical trial randomized to	Responder vs non-responder ≥50% reduction in QIDS-C total score from baseline to week 8 Remission vs non-remission <5 on QIDS-C	Linear penalized Regression XGBoost *Best performance using linear penalized regression*	153 metabolites within five analyte groups: acylcarnitines, amino acids, biogenic amines, glycerophospholipids, and sphingolipids Six functionally validated pharmacogenomic SNP biomarkers in or near the TSPAN5,	Top predictors varied by algorithm and feature set, but hydroxylated sphingomyelins, glycerophospholipids, clinical/sociodemographic features, and acylcarnitines, and were all represented.	Citalopram / Escitalopram 75.3% Citalopram/ Escitalopram/Placebo 72.7-76.6%	SSRI models: (Metabolomics alone) Linear penalized regression - AUC: 0.84 XGBoost - AUC: 0.75 (metabolomics + SNPs)

		either: 1. escitalopram (≤ 20/mg day) + placebo 2. bupropion (≤ 400 mg/day) + escitalopram 3. extended-release venlafaxine (≤ 300 mg) + mirtazapine (≤ 45 mg/day)			ERICH3, DEFB1, and AHR genes			Linear penalized regression - AUC: 0.86 XGBoost - AUC: 0.74 SSRI + placebo models: (Metabolomics alone) Linear penalized regression - AUC: 0.85 XGBoost - AUC: 0.75 (metabolomics + SNPs) Linear penalized regression - AUC: 0.86)

								XGBoost - AUC: 0.83
Nguyen, 2022	222 patients with MDD enrolled in the EMBARC trial	Trial contained two 8-week stages: First randomized in a double-blind manner to sertraline or placebo arms. At week 8, patients who did not meet response criteria (Clinical Global improvement score less than "much improved" were crossed over under double-blind conditions to bupropion treatment.	*Classification and regression models* Change in HAMD over 8-week treatment stage (week 8 minus baseline for sertraline and placebo, week 16 minus week 8 for bupropion) Responders vs. Non-responders ≥50% reduction in HAMD from pretreatment	Feed-forward neural networks *Data augmentation, a process used in deep learning to reduce the likelihood of overfitting, was used, which generates additional image data by causing slight distortion to the original acquired images.*	Contrast maps parcellated into 200 functional brain regions during number-guessing trial, reward expectancy and prediction error 95 pretreatment clinical measures and demographic features acquired on the same day as imaging	*Top 20 predictors (Sertraline):* SCID Psychomotor agitation 17-item HAMD total 24-item HAMD total PE Cerebellum Crus1 L2 FHS family hx. suicide SCQ total SCID Age of first dysphoria RE SupraMarginal R 2 RE Frontal Inf Tri R AN Frontal Sup L SCID Age of first MDD episode RE SupraMarginal R 2 RE Frontal Inf Tri R AN Frontal Sup L	Sertraline: R^2=0.48 RMSE=5.15 Placebo: R^2=0.28 RMSE=5.87 Bupropion: R^2=0.34 RMSE=4.46	*Remissio*n Sertraline: AUC: 0.60 PPV: 0.69 Placebo: AUC: 0.65 PPV: 0.81 Bupropion: AUC: 0.71 PPV: 0.75 *Response* Sertraline: AUC: 0.62 PPV: 0.68

			Remission vs non-remission HAMD ≤7 at week 8.			SCID Age of first MDD episode AN Occipital Mid R RE Cingulum Post R Employed part-time FHS family hx. Mania PE Temporal Sup R 2 FHS family hx. hallucinations NEO Conscientiousness score Unemployed AN Temporal Mid R *Top 20 predictors (Placebo)* SCID Current panic disorder Age at evaluation SCID Hypersomnia Marital status - Separated		Placebo: AUC: 0.67 PPV: 0.69 Bupropion: AUC: 0.57 PPV: 1.00

						PE Cerebellum 4 5 L Asian race AN Occipital Mid R 2 NEO Openness score MASQ Anhedonic Depression score SCID Longest period w/o dysphoria EHI Handedness score PE Cerebellum 9 L 2 RE SupraMarginal R PE Temporal Sup L 2 STAI post-fMRI score PE Occipital Inf R PE Temporal Mid R 3 PE Parietal Sup L 2 AN Occipital Inf L PE Frontal Mid L *Top 20 predictors*		

						(Bupropion) Years of education Highest education level - high school AN Cerebellum 9 R PE Cingulum Mid R RE Caudate L FHS family hx. mental illness RE Frontal Mid Orb L 2 SCID current episode anxious distress FHS family hx. Depression PE Caudate R AN Cingulum Post R 2 RE Cerebellum Crus1 R AN Frontal Sup R PE Lingual R RE Hippocampus R 2		

						RE Cerebellum Crus1 L 2 MASQ Anxious Arousal score CHRTP propensity score PE Lingual L AN Vermis 10		

Rajpurkar, 2020	518 patients with MDD	Patients randomized in a 1:1:1: ratio to escitalopram, sertraline, or extended-release venlafaxine for 8 weeks (iSPOT-D)	Predicting the Improvement for Each Symptom of the HRSD-21 Depression Assessment Scale	GBM	Baseline symptoms and pre-treatment EEG data	Important EEG features included: O1 alpha absolute (3.0% - Physical Anxiety) T7-T3 alpha absolute ratio (6.7%-Trouble Sleeping) T7-T3 beta absolute ratio (4.4% - Trouble sleeping) F7 gamma relative (5.1% - Weight loss) Fp2 delta relative (4.4% - Weight loss) F8 theta relative (2.9% - Agitation) F3 alpha absolute	R2 0.375-0.551 *Best model observed using EEG and baseline symptom features*	95% CI: 0.473-0.639 Used C-index to assess performance (probability that the algorithm will correctly identify, given 2 random patients with different improvement levels, which patient showed greater improvement

						(2.4% - Appetite change) Fp2 theta absolute (2.4% - Appetite change) T7-T3 beta relative ratio (4.7% - Unreality and nihilism) F7 beta relative (3.3% - Unreality and nihilism)		
Rethorst, 2017	122 patients with MDD	Patients were randomized to one of two exercise dose groups for 12 weeks: 4 or 16 kcal/kg/week (TREAD trial)	Patients were categorized into "remitters" (≤12 on the IDS-C), non-responders (<30% drop in IDS-	LASSO RF	25 clinical variables, five baseline serum biomarkers (IL-1B, IL6, TNF-α, SHAPS, BDNF)	*Remission* BDNF PANAS (positive) IDS-SR IL-β	NA	AUC (average from both models) Remission: 0.785 Nonresponse: 0.710

			C), or neither.			*Non-responder* VO2max PANAS (positive) BDNF IL-6		
Taliaz, 2021	1829 patients with MDD Training set: 1167 patients (STAR*D) Testing set: 271 patients (STAR*D) Validation set: 259 patients (STAR*D) External validation set:	STAR*D Largest prospective clinical trial of major depressive disorder ever conducted; comprised 4 levels of treatment according to clinical response. PGRN-AMPS 8-week clinical trial randomized to either: 1. escitalopram (10/mg day) 2. citalopram	Responders vs Non-responders 1) exponential response Continuous measure representing median antidepressant improvement rates for each of the STAR*D treatments, and used to partition patients into responders/	SVM *linear kernel*	Final model comprised 43 features (27 genetic variants, 9 clinical features, and 7 demographic features) Genetic components comprised brain-related terms (40%), neuronal signalling-related terms (40%), and 20% comprised other terms (e.g., regulation of body fluid levels)	<u>Citalopram model:</u> OPRM1 ZFPM2 WWOX Depression severity Employment Age Marital Status Education <u>Venlafaxine model:</u> STK39 CERS6 CCDC63	STAR*D model (validation set) <u>Citalopram</u> 60.5% <u>Venlafaxine</u> 74.3% <u>Sertraline</u> 75.5% PGRN-AMPS model (External validation set)	STAR*D model (Validation set) <u>Citalopram</u> Sensitivity: 67% Specificity: 54% PPV: 59.3% NPV: 62% <u>Venlafaxine</u> Sensitivity: 70% Specificity: 78.6% PPV: 76.6% NPV: 72.4%

	132 patients (PGRN-AMPS)	(20/mg day)	non-responders 2) Classic response ≥ 50% reduction in QIDS from baseline to each treatment 85% agreement in response between two definitions, however there was a discrepancy in 15% of cases			<u>Sertraline model:</u> MTOR HS6ST3 PRKCA GRIA1 GRIN2A IFNA1 FKBP5 GRIK4 Anxiety Disorders Neurological system problems Musculoskeletal/ Integumentary system problems History of medication use Employment Residence Age	<u>Citalopram</u> 61.3%	<u>Sertraline</u> Sensitivity: 69.2% Specificity: 81.8% PPV: 79.2% NPV: 72.7% PGRN-AMPS model (External validation set) <u>Citalopram</u> Sensitivity: 75.5% Specificity: 47.1% PPV: 58.8% NPV: 65.8%

Table 1 – Data-driven Biomarkers and Model Performance

Abbreviations:

AA, Arachidonic Acid; *AAP*, Atypical Antipsychotics; *ALA*, α-Linolenic Acid; *BPRS*, Brief Psychiatric Rating Scale; *AN*, anticipation; *CAPS*, Clinician Administered PTSD Scale; *CBM*, Connectome Based Predictive Model; *CBT*, Cognitive Behavioural Therapy; *CFP*, Cingulo-frontal-parietal cognitive/attention network; *CI*, Confidence Interval; *CO-MED*, Combining medications to enhance depression outcomes; *CPM*, Connectome-based predictive modeling; *DA*, Discriminant Analysis; *DF*, Deterministic Forest; *DHA*, Docosahexaenoic Acid; *DPA*, Docosapentaenoic Acid; *DT*, Decision Tree; *ELM, Extreme Learning Machine; EPA*, Eicosapentaenoic Acid; *FDR, Fisher Discriminant Ratio; FDG-PET*, (18F)Fluorodeoxyglucose PET*; FHS*, Family History Screen; *fMRI*, Functional Magnetic Resonance Imaging*; GBM, Gradient Boosting Machine; GM, Gray Matter; GPC, Gaussian Process Classification; Hx*, History; *ICA*, Independent Component Analysis*; IDS*, Inventory of Depressive Symptomatology; ISPC ,International SSRI Pharmacogenomics Consortium; iSPOT-D, international Study to Predict Optimized Treatment in Depression; IT, Interaction Tree; *LASSO, Least Absolute Shrinkage and Selection Operator; LDA, Linear Discriminant Analysis; LGEM, Latent Group Effectiveness Modeling; LR, Logistic Regression; MDD, Major Depressive Disorder;* Mid, middle; *MRF, Midline Right Frontal; NEO-FFI*, NEO-Five Factor Inventory; *NPV, Negative Predictive Value; OCD, Obsessive Compulsive Disorder; PANSS-NS, Positive and Negative Symptom Scale (negative scale); PCA, Principal Component Analysis; PE*, Prediction Error; *PF, Prefrontal; PGRN-AMPS*, Pharmacogenomics Research Network Antidepressant Medication Pharmacogenomic Study*; PHDD, Percentage of Heavy Drinking Days; PHQ, Patient Health Questionnaire; PPV, Predictive Positive Value; PUFAs, Polyunsaturated Fatty Acids; SCID*, Structured Clinical Interview for DSM-5; *SCQ*, Self-Administered Comorbidity Questionnaire; *SELSER, Sparse EEG Latent SpacE Regression; sMRI, Structural Magnetic Resonance Imaging; SNP, Single-Nucleotide Polymorphisms; SNRI, Serotonin and Norepinephrine Reuptake Inhibitor; SPECT*, Single-photon emission computerized tomography; *SSRI, Selective Serotonin Reuptake Inhibitor; Sup*, Superior; *SVM, Support Vector Machine; SVR, Support Vector Regression; RE*, Reward expectancy; *RF, Random Forest; rs-fMRI, Resting State Functional Magnetic Resonance Imaging; RVM, Relevance Vector Machine; tDCS*, Transcranial Direct Current Stimulation; TRD, Treatment-Resistant Depression; TREAD, TREAting Depression with physical activity; UHR, Ultra-High Risk; VT, Virtual Twins; *ω-3, Long-chain Omega-3; WM, White Matter; XGBoost*, Extreme Gradient Boosting

[1]All studies used DSM-IV criteria for diagnosis, except when specified otherwise. [2]The sample size showed in the table includes only the number of patients used for the model development, and does not include healthy controls used for other purposes

Authors	Classification Task	Method to address class imbalance	True and False Positives/Negatives	Performance Metrics	95% Confidence Intervals of Accuracy
PERIPHERAL BLOOD MARKERS					
Amminger, 2015	*Responders* (≥ 15-point increase in the GAF) vs *Non-responders* (22/18)	N/A	TP = 19 FP = 2 TN = 16 FN = 3	Balanced Accuracy = 86.7% Sensitivity = 86.7% Specificity = 86.7%	Accuracy = 86.70% (95% CI: 73.20-95.81)
Maciukiewicz, 2018	*Responders* (≥50% improvement in MADRS) vs *Non-responders* (27/11) *Remission* (<10 MADRS) *vs.* *Non-remission* (19/18)	N/A	*Response* TP = 23 FP = 10 TN = 1 FN = 4 *Remission* TP = 13 FP = 10 TN = 8 FN = 6	*Response* Balanced Accuracy = 47% Sensitivity = 87% Specificity = 7% *Remission* Balanced Accuracy = 57% Sensitivity = 68% Specificity = 46%	*Response* Accuracy = 63.12% (95% CI: 45.95-78.15) *Remission* Accuracy = 56.74% (95% CI: 39.48-72.89)
ELECTROENCEPHALOGRAPHY					
Al-Kaysi, 2017	*Responders* (≥50% improvement in MADRS) vs *Non-responders* (5/5)	N/A	N/A	Accuracy = 76%	N/A
Cao, 2019	*Responders* (≥45% improvement in HAMD-17) vs *Non-responders*	Oversampling minority class	TP = 13 FP = 2 TN = 19 FN = 3	Balanced Accuracy = 87% Sensitivity = 82.1% Specificity = 91.9%	Accuracy = 81.3% (95% CI: 71.23-95.47)

	(16/21)				
de la Salle, 2020	*Responders* (≥50% improvement in MADRS) vs *Non-responders* (27/20)	N/A	TP = 19 FP = 1 TN = 19 FN = 8	Balanced Accuracy = 82.5% Sensitivity = 70% Specificity = 95%	Accuracy = 80.96% (95% CI: (66.87-90.93)
Jaworska, 2019	*Responders* (≥50% improvement in MADRS) vs *Non-responders* (27/24)	N/A	TP = 21 FP = 0 TN = 24 FN = 6	Balanced Accuracy = 88% Sensitivity = 77% Specificity = 99%	Accuracy = 88.24% (95% CI: 76.14-95.56)
NEUROIMAGING					
Braund, 2022	*Responders* (≥50% improvement in $HAMD_{17}$ or $QIDS\text{-}SR_{16}$) vs *Non-responders* (102/127)	N/A	TP = 64 FP = 19 TN = 107 FN = 39	Balanced Accuracy = 73.75% Sensitivity = 62.5% Specificity = 85.0%	Accuracy = 75% (95% CI: 57.8-87.9)
Klöbl, 2020	*Responders* (≥50% improvement in $HAMD_{17}$) vs *Non-responders* (19/10) *Remission* ($HAMD_{17} \leq 7$) vs *Non-remission* (16/13)	N/A	N/A	AUC = 0.68-0.73 *Sensitivity and specificity not reported*	N/A
Koutsouleris, 2017	*Good clinical response* (≥60 GAF at endpoint) vs	N/A	*Full sample* TP = 81 FP = 29	*Full sample* Balanced Accuracy = 72.1%	Accuracy = 74.56% (95% CI: 69.52-79.15)

	Poor clinical response (110/224)		TN = 156 FN = 67 *Independent sample of 108 patients* TP = 13 FP = 5 TN = 64 FN = 5	Sensitivity = 69.6% Specificity = 74.5% *Independent sample* Balanced Accuracy = 71.7% Sensitivity = 71.1% Specificity = 72.2%	Accuracy = 82.49% (95% CI: 72.85-89.80)
Nord, 2019	*Responders* (≥50% improvement in $HAMD_{17}$) vs *Non-responders* (10/10)	N/A	N/A	AUC = 0.856 *Sensitivity and specificity not reported*	N/A
Sarpal, 2016	*Responders* (Two consecutive visits with CGI improvement score of 1-2 and ≤3 in conceptual disorganization, grandiosity, hallucinatory behavior, and unusual thoughts on the BPRS-A) vs Non-responders (44/37)	N/A	*Independent sample of 40 patients* TP = 16 FP = 5 TN = 15 FN = 4	Balanced Accuracy = 77.5% Sensitivity = 80% Specificity = 75%	Accuracy = 77.50% (95% CI: 61.55-89.16)
Yip, 2019	*Abstinence* (Cocaine negative urine) vs. *relapse* (Cocaine positive urine) *Exact numbers of each class not reported*	N/A	N/A	Sensitivity = 35% Specificity = 82%	N/A

MULTIMODAL DATA					
Ambrosen, 2020	*Responders* (20% decrease in PANSS) vs. *Non-responders* (71/68)	N/A	N/A	N/A	N/A
Athreya, 2019	*Responders* (≥50% improvement in $HAMD_{17}$ or QIDS-C) vs. *Non-responders* (379/253) (n=467 STAR*D; n=165 ISPC)	N/A	*$HAMD_{17}$ – men* TP = 188 FP = 52 TN = 134 FN = 92 *$HAMD_{17}$ – women* TP = 191 FP = 69 TN = 118 FN = 90 *QIDS-C – men* TP = 79 FP = 19 TN = 47 FN = 20 *QIDS-C – women* TP = 77 FP = 21 TN = 45 FN = 22	*$HAMD_{17}$ – men* Balanced Accuracy = 69.5% Sensitivity = 67% Specificity = 72% *$HAMD_{17}$ – women* Balanced Accuracy = 65.5% Sensitivity = 68% Specificity = 63% *QIDS-C – men* Balanced Accuracy = 75.5% Sensitivity = 80% Specificity = 71% *QIDS-C – women* Balanced Accuracy = 73% Sensitivity = 78% Specificity = 68%	*$HAMD_{17}$ – men* Accuracy = 69.10% (95% CI: 64.69-73.27) *$HAMD_{17}$ – women* Accuracy = 66.02% (95% CI: 61.53-70.31) *QIDS-C – men* Accuracy = 76.36% (95% CI: 69.14-82.92) *QIDS-C – women* Accuracy = 73.94% (95% CI: 66.54-80.45)
Crane, 2017	*Responders* (≥50% improvement in $HAMD_{17}$) vs.	N/A	N/A	Balanced Accuracy = 82.1% Sensitivity = 84.2%	N/A

	Non-responders (Exact number of responders/non-responders was not reported)			Specificity = 80.0%	
Fonzo, 2017	*Remission* (≤ 20 post-treatment CAPS score) vs. *Non-remission* (Exact number of responders/non-responders was not reported)	N/A	N/A	Accuracy = 79.5-97.7% *Sensitivity and specificity not reported*	N/A
Lee, 2018	*Good outcome* (Decrease in PANSS) *vs.* *Poor outcome* (Olanzapine – 85/23; Ziprasidone – 24/27)	N/A	*Olanzapine* TP = 69 FP = 11 TN = 16 FN = 12 *Ziprasidone* TP = 33 FP = 5 TN = 6 FN = 8	*Olanzapine* Balanced Accuracy = 66.5% Sensitivity = 81% Specificity = 52% *Ziprasidone* Balanced Accuracy = 74.5% Sensitivity = 75% Specificity = 74%	*Olanzapine* Accuracy = 79.66% (95% CI: 70.84-86.80) *Ziprasidone* Accuracy = 74.96% (95% CI: 61.01-85.94)
Joyce, 2021	*Responders* (≥50% improvement in $HAMD_{17}$) vs. *Non-responders* (Model 1 - 48/29; Model 2 – 45/26)	Oversampling minority class	*Model 1 - metabolomic* TP = 33 FP = 3 TN = 26 FN = 15 *Model 2 - multi-omics* TP = 32 FP = 3 TN = 23 FN = 13	*Model 1 - metabolomic* Balanced Accuracy = Sensitivity = 69% Specificity = 90% *Model 2 - multi-omics* Balanced Accuracy = Sensitivity = 71% Specificity = 88%	*Model 1 - metabolomic* Accuracy = 76.63% (95% CI: 65.60-85.52) *Model 2 - multi-omics* Accuracy = 77.48% (95% CI: 66.02-86.55)

Rethorst, 2017	*Remitters* (≤12 on the IDS-C) *Non-responders* (<30% improvement in IDS-C) vs. *Responders* (neither) (36/56/30)	N/A	N/A	Response – AUC = 0.785 Non-response – AUC = 0.710	N/A
Taliaz, 2021	*Exponential responders* (Continuous measure representing exponential fit for individual longitudinal measurements of QIDS during a specific treatment) vs. *Non-responders* *Venlafaxine* 41.7% response (R = 10; NR = 14) *Sertraline* 41.7% response (R = 10; NR = 14) *Citalopram* 44.6% response (R = 112; NR = 139)	N/A	*Venlafaxine* TP = 7 FP = 3 TN = 11 FN = 3 *Sertraline* TP = 9 FP = 2 TN = 4 FN = 9 *Citalopram* TP = 75 FP = 64 TN = 75 FN = 37	*Venlafaxine* Balanced Accuracy = 74.3% Sensitivity = 71% Specificity = 88% *Sertraline* Balanced Accuracy = 75.5% Sensitivity = 69.2% Specificity = 81.8% *Citalopram* Balanced Accuracy = 60.5% Sensitivity = 67% Specificity = 54%	*Venlafaxine* Accuracy = 75.00% (95% CI: 53.29-90.23) *Sertraline* Accuracy = 78.97% (95% CI: 57.09-92.97) *Citalopram* Accuracy = 59.76% (95% CI: 53.41-65.88)

Table 2: 95% Confidence Intervals of Clinical Response

Performance metrics across predictive models of treatment response within randomized clinical trials. In cases where confidence intervals were not reported, this metric was calculated using true/false positives and negatives, as well as the prevalence of responders within each study. In instances where true/false

positives and negatives were not reported, this was imputed using the sensitivity, specificity, and prevalence of studies. Studies that did not report these prerequisite summary statistics are indicated with N/A.

Abbreviations: *BPRS-A*, Brief Psychiatric Rating Scale-Anchored; *GAF*, Global Assessment of Functioning, *$HAMD_{17}$*, Hamilton Depression Rating Scale 17-item, *IDS*, Inventory for Depressive Symptomatology; *ISPC*, International SSRI Pharmacogenomics Consortium; *MADRS*, Montgomery-Asberg Depression Rating Scale; *PANSS*, Positive and Negative Syndrome Scale; *PGRN-AMPS*, Pharmacogenomics Research Network Antidepressant Pharmacogenomics Study; *$QIDS\text{-}SR_{16}$*, Quick Inventory of Depressive Symptomatology, *STAR*D*, Sequenced Treatment Alternatives to Relieve Depression

Supplementary Table S1 – Machine learning studies predicting treatment response in psychiatric disorders (non-randomized open-label trials)

First author, year	Sample size and diagnosis[1,2]	Open-label Trial	Outcome	Machine learning model	Data utilized	Top Data-Driven Biomarkers (Features)	Accuracy	Additional Performance Metrics
STUDIES USING ELECTROENCEPHALOGRAPHIC MEASURES								
Arns, 2012	90 patients with treatment-resistant depression	Average of 20 sessions of left DLPFC 10 Hz rTMS treatment	Responders vs Nonresponders ≥ 50% reduction in BDI from baseline to post-treatment	LDA	*resting-state EEG* EEG theta power Alpha peak frequency PF Delta Cordance	Anterior iAPF, P300 amplitude at Pz, prefrontal delta and beta cordance	NA	AUC: 0.814
Bruder, 2008	18 patients with MDD	12-weeks of open-label fluoxetine	Responders vs Nonresponders CGI-I rating of "much or very much" improved	LDA	*resting-state EEG* Alpha power and asymmetry at	Alpha power and asymmetry at occipital sites	NA	PPV: 72.7-77.8% NPV: 55.6-80.0%

			considered as responders		occipital sites			
Erguzel, 2016	147 patients with treatment-resistant depression	3 weeks (20 sessions) of open label adjunctive rTMS	Responders vs Nonresponders ≥ 50% reduction in HAM-D from baseline to post-treatment	ANN SVM DT	*resting-state EEG* Cordance (combination of absolute and relative power)	Not available	78.3-86.4% *Best performance using SVM*	SVM AUC: 0.918 ANN AUC: 0.877 DT AUC: 0.807
Hasanzadeh, 2019	46 patients with MDD	5-sessions of left DLPFC 10 Hz rTMS treatment	Responders vs Nonresponders ≥ 50% reduction in BDI-II or HRSD scores Remission vs non-remission	kNN	*resting-state EEG* Nonlinear features (LZC, CD, KDF) Power spectrum features (delta,	Power (D,T, A, B)- 91.3% *Composite measures* All - 87% Bispectrum - 84.8% Nonlinear (LZC, KFD, CD) - 80.4%	78.3-82.6% *best performance with Lempel-Ziv complexity feature extraction*	Sensitivity: 78.3-82.6% Specificity: 73.9-91.3A%

			BDI ≤ 8		theta, alpha, beta) Bispectrum features (2D Fourier transform of the third order cumulant)	Cordance - 76.1% *Single measures* Power-B - 91.3% BisplSL-D - 89.1% BisplSL-B - 87% Bisp2M-D - 84.8% BispEn-D - 82.6%		
Salle, 2020 *	47 patients with MDD	12-week double-blind trial of: 1) escitalopram 2) bupropion 3) escitalopram + bupropion	Responders vs Nonresponders ≥50% MADRS score reduction from baseline Remitters vs Nonremitters ≤10 MADRS at 12 weeks	LR	*Pre-treatment rs-EEG* Baseline PF and MRF Theta Cordance	Response: Change in PF Cordance ≤-0.81 Change in MRF ≤0.02 Remission: Change in PF Cordance ≤-0.81 Change in MRF ≤ 0.54	74-81% *Response* 51-70% *Remission* *Best performance in both models using PF Cordance alone*	Response - PF Cordance AUC: 0.85 Sensitivity: 70% Specificity: 95% PPV: 0.95 NPV: 0.64 Remission - PF Cordance AUC: 0.66 Sensitivity: 0.70

								Specificity:0.63 PPV: 0.58 NPV: 0.74
Zandvakili, 2019	29 patients with comorbid MDD and PTSD	Unblinded trial of 5 Hz rTMS to the left DLPFC (F3) for three weeks	Responders vs Nonresponders ≥ 50% reduction in IDS-SR	LASSO SVM	*resting-state EEG* EEG coherence (alpha, beta, theta, delta)	Alpha band, local left prefrontal connections (contributed 12.13% accuracy,95% CI: 9.18%−14.85% on bootstrap) Prefrontal electrodes and midline electrodes contributed 7.26% (95% CI: 4.31%−9.86%), but performance did not depend on local-midline connections.	75.4-78.4%	MDD AUC: 0.83 Sensitivity: 47-94% Specificity: 0-83% PTSD: AUC: 0.71 Sensitivity: 37-100% Specificity: 0-100%
Zhdanov, 2020	122 patients with MDD	Multicentre open-label trial of	Responders vs Nonresponders	SVM *Radial kernel*	*resting-state EEG (baseline + week 2)*	High alpha-band-power in anterior cingulate cortex was the most prominent	79.2% *Using baseline*	*Baseline Model* Sensitivity - 67.3%

		escitalopram (10-20mg) treatment	≥ 50% reduction in MADRS from baseline to week 8		Electrode-level frequency analysis power spectral features in the source domain spatiotemporal complexity global brain network dynamics	predictive feature shared by all the feature sources. High-alpha-band power in rostral anterior cingulate cortex appeared in baseline and week 2 data and high-beta-band at week 2 only	*EEG data* *82.4%* *Using baseline and week 2 EEG data*	Specificity - 91.0% *Baseline and Week 2 Model* Sensitivity: 79.2% Specificity: 85.5%
STUDIES USING NEUROIMAGING								
Ananth, 2020	27 patients with bipolar depression	8 weeks of lithium monotherapy titrated to a therapeutic plasma level of 0.8-1.2	Responders vs Nonresponders ≥ 50% reduction in	LASSO	PET 5-HTT and 5-HT1A binding	Amygdala, hippocampus, and parahippocampal gyrus were found to be important features, with all other features shrunk to zero.	87.7%	87.5% sensitivity 80% specificity

		mEq/l	HDRS-24 pre to post-treatment		12 ROIs			
Brown, 2020	20 patients with either MDD or Bipolar Depression	DBS implanted into SCC white matter bilaterally	Responders vs Nonresponders ≥ 48% reduction in HDRS scores from baseline to 6 months postoperatively	Gaussian NB	Baseline mean FDG-PET signal intensity	Baseline mean FDG-PET signal intensity from the SCC ROI could predict which patients responded to treatment with an accuracy of 80%.	80%	Sensitivity: 80% Specificity: 80%
Cao, 2018	43 drug-naive inpatients with first-episode schizophrenia	10-week open-label risperidone treatment	Responders vs Nonresponders ≥ 30% reduction in PANSS total score	SVM *linear*	rs fMRI	Left fusiform - ***t***=4.55 Right precentral cortex - ***t***=4.26 Right cuneus cortex - ***t***= 4.01 left fusiform - ***t***=4.87 left lingual - ***t***=4.15 Right postcentral cortex - ***t***=4.04 Right fusiform - ***t***=4.04	82.5%	Sensitivity: 88.0% Specificity: 76.9%

						Left lingual - ***t***=4.08		
Cao, 2018b	24 inpatients with MDD	8-sessions of ECT	Remission vs Non-remission Post-treatment HAM-D total score ≤ 7	SVM *linear*	sMRI Hippocampal subfield volumes	Significant volume increases in bilateral GCL and right CA3, CA4, molecular layer, and subiculum in remitters	83.3%	Sensitivity: 91.7% Specificity: 75% AUC: 0.90
Cash, 2019	47 patients with MDD	5-8 weeks of rTMS treatment targeting region F3	Responders vs Nonresponders > 25% change in MADRS scores	SVM	rs fMRI Voxel-wise BOLD signal power resting state network connectivity	Lower BOLD power in caudate, prefrontal cortex, and thalamus, as well as FC in the DMN and affective networks were associated with treatment response	85%	92% specificity 75% sensitivity
Ge, 2020	32 patients with treatment-resistant MDD	20-30 sessions of 10 Hz (high-frequency left stimulation) or intermittent theta-burst	Responders vs Nonresponders ≥ 50% reduction in HRSD from baseline	LDA	rACC-IPL and sgACC-DLPFC based FC	Stronger the FC between rACC and IPL, greater improvement on HRSD (r=0.49, p=3.48 x 10-4) Stronger the FC between sgACC and right DLPFC,	76-84% *best performance using rACC-IPL features*	rACC-IPL 84% (sensitivity: 81%, specificity: 86%) sgACC-DLPFC

		rTMS over the left DLPFC				lesser improvement on HRSD (r=-0.62, p=1.95 x 10-6)		76% (sensitivity: 48%, specificity: 97%)
Gong, 2020	57 patients with SCZ	12-sessions of ECT in conjunction with standard antipsychotic drugs	Regression model - continuous improvement in symptoms (PANSS)	SVR	sMRI dMRI GM tissues of 23 ROIs and the FA values of 37 WM tracts	Calcarine_L-Temporal_Pole_Sup_L Lingual_R-Temporal_Mid_R Occipital_Mid_L-Temporal_Inf_L Frontal_Inf_Orb_R-Insula_R Frontal_Inf_Orb_R-Insula_R Occipital_Mid_R-Temporal_Mid_R	N/A	RMSE: 14.980
Hahn, 2015	49 medication-free patients with PD/AG	12-sessions of CBT	Responders vs Nonresponders > 50% reduction in	GPC	fMRI during a differential fear-conditioning task	*Top 10% whole-brain GPC weights* Precentral gyrus - 3.19	82%	Sensitivity: 92% Specificity: 72%

			HARS scores			Occipital fusiform gyrus - 3.04 Frontal orbital cortex - 2.79 Middle temporal gyrus (temporo-occipital part) - 2.78 Putamen - 2.68 Supramarginal gyrus (anterior division) - 2.47 Frontal pole - 2.23 Occipital pole - 2.15 Inferior frontal gyrus (pars triangularis) - 2.15 Postcentral gyrus - 2.03		
Leaver, 2018	46 patients with MDD	Right-unilateral ECT	Responders vs Nonresponders ≥ 50% reduction in composite depression scores	SVM *radial kernel*	sMRI arterial spin-labeled-fMRI	*Most significant features in responders* Left thalamus - p, RFT corrected = 2.50 x 10-6 Left somatomotor cortex - p, RFT corrected = 3.68 x 10-5 Left occipital cortex - p, RFT corrected =.0438	58-68%	Sensitivity: 54-64% Specificity: 55-74%

						Right angular gyrus = 1.09 x 10-10 Right frontal operculum = .00622 Precuneus = .031 *Most significant features in nonresponders* Right hippocampus and accumbens = 1.79 x 10-7 Posterior cingulate cortex = .000767		
Månsson, 2015	26 patients with SAD	Open-label cross-over trial: 1) 9-week guided internet CBT 2) ABM	Responders vs Nonresponders Post-treatment scores (1 year follow-up) of 1-2 on the CGI-I as responders ≥ 3 on post-treatment CGI-I classified as	SVM *linear*	fMRI during a self-referential criticism task	ACC - 91.7% Amygdala - 47.7% dlPFC - 43.2% Hippocampus - 51.9% Insula - 43.6% vmPFC - 39.0%	39-91.7% *Best performance observed using ACC*	Sensitivity: 41.7-83.3% Specificity: 36.4-100% AUC: 0.29-0.91

			nonresponders					
Wade, 2016	53 patients with MDD	4-6 weeks of open-label ECT (3 treatments per week)	Responders vs Nonresponders Response defined as > 50% improvement in HAM-D scores over the course of treatment.	SVM *radial*	Siemens 3T Allegra (T1-Structural MRI) radial distance and Jacobian determinant in the accumbens, caudate, putamen and pallidum	Significant volumetric gain in the accumbens F(2, 18.98)=9.18, P=0.002, in responders	72%	AUC: 0.54 (95 % CI=29-78%)
Xi, 2020	57 patients with SCZ	9-12 sessions of unilateral ECT (800 mA stimulus intensity)	Responders vs Nonresponders ≥ 70% reduction in PANSS total scores	SVM	GE Discovery MR750 3T (T1-structural MRI) GM volume in 19 ROIs (258 features)	Top features included cortical (inferior frontal gyrus, cingulate cortex, and temporal and parietal lobes) and subcortical regions (insula, thalamus, and hippocampus)	87.59%	N/A
STUDIES USING MULTIMODAL DATA								
Bailey, 2018	57 patients with treatment-	5-8 weeks of rTMS	Responders vs Nonresponders	SVM *linear*	Pre-treatment resting-state EEG and mood	Responders showed more theta connectivity relative to non-responders (p=0.0216,	86.60%	Sensitivity: 84% Specificity: 89%

	resistant MDD		Response defined as > 50% improvement in HDRS scores		symptoms	FDR p=0.030)		
Ball, 2014	48 adults 25 patients with GAD 23 patients with PD	10-sessions of open-label weekly individual CBT	Responders vs Nonresponders OASIS scores of ⩽5 at the end of therapy	RF	Clinical/ socio-demographic data, and an fMRI task appraising emotional responses to negative images reappraise- and maintain-related activation before treatment in each of the 70 anatomical ROIs	*Ten variables met inclusion in final model* OASIS, ASI, PSWQ-A, as well as right hippocampus and left uncus activation during maintenance, and left transverse temporal gyrus, left supramarginal gyrus, left precentral gyrus, left superior frontal gyrus, and right substantia nigra activation during reappraisal	69-79% *Best performance observed with fMRI features alone*	Sensitivity: 79-86% Specificity: 53-68%
Luo, 2014	24 patients with cocaine dependence	12-weeks of contingency management therapy 24 sessions	Responders vs Nonresponders ≥ 1 month of abstinence (urine	LR SVM *radial kernel*	Baseline demographic variables, and striatal PET (ECAT EXACT HR+) data	Best performance using change in binding potential in the ventral striatum and posterior caudate at week 2, 3 and 4	82-96% *Best performance observed using neuroimagi*	N/A

		total (Community Reinforcement Approach)	measurements)				*ng and behavioral predictors*	
Kim, 2015	83 patients with ADHD	8-week open label trial of methylphenidate	Responders vs Nonresponders Post-treatment scores (8-weeks) of 1-2 on the CGI-I as responders ≥ 3 on post-treatment CGI-I classified as nonresponders	SVM *2nd order polynomial kernel* DT RF LRR	Genomic DNA extracted from whole blood lymphocytes using a G-DEXTM II polymorphism and 40-base pair VNTR polymorphism located in the 3'--UTR of DAT1 were genotyped Resting-state fMRI (3T Siemens scanner) repetition time 3000 ms; echo time 40 ms; acquisition matrix 128× 128; field of	Wrapper subset evaluation method demonstrated the age, weight, ADRA2A MspI and Dra I polymorphisms, lead level, SCWT color-word and word performance, and oppositional symptoms of DBD as the most differentiating subset of features.	(range) SVM: 64.1-84.6% DT: 61.5-69.2% RF: 61.5-73.1% LRR: 65.4-76.9%	(range) SVM: 64.1-84.6% DT: 61.5-69.2% RF: 61.5-73.1% LRR: 65.4-76.9%

					view 240× 240 mm2; flip angle 90°; voxel size 1.9 mm × 1.9 mm × 4.0 mm; slices 30.			
Martinuzzi, 2019	325 patients with first-episode psychosis OPTiMiSE Study (7 general hospitals and clinics in 14 European countries, Israel and Australia)	4 weeks of open label Amisulpride (≤ 800 mg/day)	Remission vs non-remission ≤3 on 8 PANSS items: P1, P2, P3, N1, N4, N6, G5, and G9	Sparse *k*-means (used to derive 4 patient subtypes) Regularized LR	Data acquired on the V-PLEX Sector Imager 2400 plate reader and analyzed using the Discovery Workbench 3.0 software (MSD) Proinflammatory Panel 1, Cytokine Panel 1, Chemokine Panel 1, Th17 Panel 1 and Vascular Injury Panel 2 v-PLEX®kits (MSD)	Lower serum levels of IL-15, higher serum levels of CXCL12, seropositivity to CMV, use of recreational drugs, and being younger were all associated with increased odds of being non-remitters in CA patients.	64%	AUC: 0.73 Sensitivity: 83% Specificity: 45%

Takamiya, 2020	27 patients with MDD	Bitemporal ECT 2-3 times per week until no improvement was seen within the last 2 sessions	Regression model - continuous improvement in symptoms (HAMD)	SVM *radial kernel*	Baseline demographic variables, and pretreatment sMRI T1-weighted images 3-T GE Signa HDxt scanner repetition time=6.9 milliseconds time to echo = 2.9 milliseconds slice thickness = 1.0 mm	Left gyrus rectus Right anterior lateral temporal lobe Left lateral occipital lobe Right cuneus Left putamen Left third ventricle HAMD item 3 (suicide) HAMD item 10 (anxiety psychic) Right inferior middle temporal gyrus Right third ventricle Right cerebellum Right superior temporal gyrus Left brainstem HAMD item 9 (agitation) Right brainstem	70.4-92.6% *Best performance observed using clinical and neuroimaging features*	Sensitivity: 95-100% Specificity: 0-71.4% PPV: 73.1-90.9% NPV: 0-100%

Supplementary Table S1 – Machine learning studies predicting treatment response in psychiatric disorders (non-randomized open-label trials)

Abbreviations:

AA, Arachidonic Acid; AAP, Atypical Antipsychotics; ABM, Attention Bias Modification; ACC, Anterior Cingulate Cortex; ADHD-RS, Attention Deficit and Hyperactivity Disorder Rating Scale; ADTree, Alternating Decision Tree; AIC, Akaike Information Criteria; ALA, α-Linolenic Acid; AN, Anorexia Nervosa; *ANN, Artificial Neural Networks; AUD, Alcohol Use Disorder; BDD, Body Dysmorphic Disorder; BDI-II, Beck Depression Inventory - Second Edition; BN, Bayesian Networks; BOLD, Blood-Oxygen Level-Dependent;* BSP, Brief Supportive Psychotherapy; BZD, Benzodiazepines; Calcarine-L, *Calcarine fissure and surrounding cortex*; CBM, Connectome Based Predictive Model; CBT, Cognitive Behavioural Therapy; DA, Discriminant Analysis; DBS, Deep Brain Stimulation; DF, Deterministic Forest; DHA, Docosahexaenoic Acid; dMRI, diffusion MRI; DPA, Docosapentaenoic Acid; DSB, Deep-Brain Stimulation; DT, Decision Tree; *ELM, Extreme Learning Machine; EMD, Empirical Mode Decompositions; ENRR, Elastic Net Regularized Regression;* EPA, Eicosapentaenoic Acid; ERP, Exposure and Response Prevention; FF-BP ANN, Feed-forward Back-propagation Artificial Neural Network; *FDR, Fisher Discriminant Ratio; (18F)Fluorodeoxyglucose PET; (FDG-PET), Feature Selection; fMRI, Functional Magnetic Resonance Imaging; FIBSER, Frequency, Intensity, and Burden of Side Effects Rating; Frontal_Inf_Orb_R, right Frontal gyrus, orbital part; GAD, Generalized Anxiety Disorder; GBM, Gradient Boosting Machine; GK, Gaussian Kernel; GCL,* granule cell layer; *GM, Gray Matter; GPC, Gaussian Process Classification; GPR, Gaussian Process Regression; GEE, Generalized Estimated Equation; GNB, Gaussian Naive Bayes; HARS, Hamilton Anxiety Rating Scale; HDRS, Hamilton Depression Rating Scale; ICA, Independent Component Analysis;* IPT-PS, Interpersonal Psychotherapy for Depression with Panic and Anxiety Symptoms; IDS, Inventory of Depressive Symptomatology; IT, Interaction Tree; *KL, Kullback-Leibler; KPLSR, Kernelized Partial Least Squares Regression;* L1-LR, L1 Regularized Logistic Regression; *LAR, Least Angle Regression; LASSO, Least Absolute Shrinkage and Selection Operator; LDA, Linear Discriminant Analysis; LGEM, Latent Group Effectiveness Modeling; Lingual_R, right lingual gyrus; LITHIA, Lithium Intelligent Agent (algorithm based on genetic algorithms and fuzzy systems); LR, Logistic Regression; LRR, Logistic Ridge Regression; LSO, Leave-site-out; LVSR, Linear Support Vector Regression; MDA, Mixture of Factor Analysis; MDD, Major Depressive Disorder; MET, Methadone; MER, Mixed Effects Regression; MFA, Mixture of Factor Analysis; MLP, Multi-Layer Perceptron; MPH, Methylphenidate; MRMR, Minimum redundancy and maximum relevance; NPV, Negative Predictive Value; NB, naive Bayes, OASIS, Overall Anxiety Severity And Impairment Scale; OCD, Obsessive Compulsive Disorder; Occipital_Mid_L, left middle occpital cortex; Occipital_Mid_R, right middle occipital cortex; PCA, Principal Component Analysis; PD, Panic disorder; PD/AG, Panic Disorder with Agoraphobia; PHDD, Percentage of Heavy Drinking Days; PHQ, Patient Health Questionnaire; PPV, Predictive Positive Value; PUFAs, Polyunsaturated Fatty Acids; SAD, Social Anxiety Disorder; SCC, Subcallosal Cingulate Cortex; SCZ, Schizophrenia; SELSER, Sparse EEG Latent SpacE Regression; SGD, Stochastic Gradient Descent; sMRI, Structural Magnetic Resonance Imaging; SNP, Single-Nucleotide Polymorphisms; SNRI, Serotonin and Norepinephrine Reuptake Inhibitor; SSRI, Selective Serotonin*

Reuptake Inhibitor; STFT, Short-time Fourier Transform; SVM, Support Vector Machine; SVM-L, Support Vector Machine with Linear Kernel; SVR, Support Vector Regression; SVM-RBF, Support Vector Machine with Radial Basis Function Kernel; SVM-RFE, Support Vector Machine Recursive Feature Elimination; SVR, Support Vector Regression; RBFS, Rank-Based Feature Selection; RBFSVR, Radial Basis Support Vector Regression; REM, Rapid Eye Movement; RMSE, Root Mean Square Error; RF, Random Forest; RFE, Recursive Feature Elimination; RR, Ridge Regression; rs-fcMRI, Resting-state Functional Connectivity Magnetic Resonance Imaging; *rs-fMRI, Resting State Functional Magnetic Resonance Imaging; RVR, Relevance Vector Regression;* tDCS; transcranial Direct Current Stimulation; Temporal_Pole_Sup_L, Temporal pole: superior temporal gyrus; Temporal_Inf_L, *left Temporal Inferior Cortex*; Temporal_Mid_R, *right middle temporal gyrus*; TRD, Treatment-Resistant Depression; TSD, Treatment-Sensitive Depression; UHR, Ultra-High Risk; VT, Virtual Twins; *ω-3, Long-chain Omega-3; WCST, Wisconsin Card-Sorting Task; WM, White Matter*

[1]All studies used DSM-IV criteria for diagnosis, except when specified otherwise.

[2]The sample size showed in the table includes only the number of subjects used for the model development, and does not include healthy controls used for other purposes

* Study lacked cross-validation metrics or training and testing sets and was therefore excluded.

PERIPHERAL BLOOD MARKERS										
Authors	Representative	Internal CV	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Amminger, 2015	No	Yes	A	B	No	No	Yes	Yes	No	5/9
Hou, 2015	No	Yes	B	B	No	No	Yes	No	No	4/9
Maciukiewicz, 2018	No	Yes	B	A	Yes	No	Yes	Yes	Yes	7/9
ELECTROENCEPHALOGRAPHY										
Authors	Representative	Internal CV	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Al-Kaysi, 2017	No	Yes	A	A	No	No	Yes	No	No	4/9
Cao, 2019	No	Yes	A	A	Yes	No	Yes	Yes	Yes	7/9
Jaworska, 2019	No	Yes	A	A	Yes	No	Yes	No	Yes	6/9
de la Salle, 2020	No	Yes	A	B	Yes	No	Yes	Yes	No	6/9
Wu, 2020	Yes	Yes	A	A	Yes	Yes	Yes	Yes	No	8/9

NEUROIMAGING										
Authors	Representative	Internal CV	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Braund, 2022	No	Yes	A	B	Yes	No	Yes	Yes	Yes	7/9
Fan, 2020	No	Yes	A	A	Yes	Yes	Yes	Yes	No	7/9
Fonzo, 2019	No	Yes	A	A	Yes	Yes	Yes	Yes	No	7/9
Klöbl, 2020	No	Yes	A	B	Yes	Yes	Yes	No	No	6/9
Koutsouleris, 2017	No	Yes	B	A	Yes	No	Yes	Yes	Yes	7/9
Nemanti, 2020	Yes	Yes	A	A	Yes	Yes	Yes	Yes	No	8/9
Nord, 2019	No	Yes	A	B	No	No	Yes	No	No	4/9
Sarpal, 2016	No	Yes	A	B	Yes	No	Yes	Yes	No	6/9
Yip, 2019	No	Yes	A	A	Yes	Yes	Yes	Yes	No	7/9
MULTIMODAL DATA										
Authors	Representative	Internal CV	Outcome	ML	Feature Selection	Class imbalance	Missing data	Performance	Testing/ Validation	Overall Score
Ambrosen, 2020	No	Yes	B	A	No	No	Yes	No	No	4/9

Athreya, 2019	Yes	Yes	A	A	Yes	No	Yes	Yes	No	7/9
Crane, 2017	No	Yes	B	B	Yes	No	Yes	Yes	No	6/9
Fonzo, 2017	No	Yes	B	B	Yes	Yes	Yes	Yes	No	7/9
Lee, 2018	No	Yes	B	B	Yes	No	Yes	Yes	No	6/9
Joyce, 2021	Yes	Yes	A	A	Yes	Yes	Yes	Yes	Yes	9/9
Nguyen, 2022	No	Yes	A	B	Yes	No	Yes	Yes	No	6/9
Rajpurkar, 2020	No	Yes	B	A	Yes	Yes	Yes	Yes	No	7/9
Rethorst, 2017	No	Yes	B	A	Yes	No	Yes	No	No	5/9
Taliaz, 2021	Yes	Yes	A	A	Yes	No	Yes	Yes	Yes	8/9

Supplementary Table 2: Quality Scores of All Studies

First author, year	**Data acquisition**	**Preprocessing / Quality Control Metrics**	**Imputation strategy**	**Feature extraction method** (*if applicable*)	**Feature selection method** (*if applicable*)
STUDIES USING PERIPHERAL BLOOD MARKERS					
Amminger, 2015	Fatty acid composition of the phosphatidylethanolamine phospholipid fraction quantified using capillary gas chromatography (7 features in total)	NA	NA	NA	NA
Hou, 2015	Genotyping of long and short alleles of the functional insertion-deletion polymorphism (5'-HTTLPR) in the promoter region of the SLC6A4 gene.	NA	NA	NA	NA

	(21 features in total)				
Maciukiewicz, 2018	Infinium PsychArray BeadChip by Illumina 571,054 SNPs were genotyped 19 SNPs retained with the highest β-coefficients	Ancestry control - Multidimensional scaling (MDS) using PLINK Control for minor allele frequency >1% Hardy-Weinberg equilibrium ($p > .0000001$) Genotype call rate (> 98%) Individual missingness (<10%)	Whole-genome IMPUTE v2.2 in 5-Mb segments per chromosome after pre-phasing with SHAPEIT2 and the 1000 genomes reference panel	NA	LASSO (non-zero β-coefficients)
First author, year	**Data utilized**	**Preprocessing / Quality Control Metrics**	**Imputation strategy**	**Feature extraction method** (*if applicable*)	**Feature selection method** (*if applicable*)
STUDIES USING ELECTROENCEPHALOGRAPHY MEASURES					
Al-Kaysi, 2017	*64-channel BrainAmp*	Downsampled to 2 KHz	NA	*Power spectral density*	NA

	MR Plus amplifiers (62 channels used) Reference electrode: Fz and Cz Sampling rate: 5 KHz	High-pass filter (Butterworth filter with 0.5 Hz cut-off frequency) Artifact removal using ICA		delta (0.5-4 Hz) theta (4-8 Hz) alpha (8-12 Hz) beta (13-30 Hz) gamma (30-100 Hz) *Alpha asymmetry* (frontal, central, and parietal cortex)	
Cao, 2019	*Mindo-4S Jellyfish* (four dry electrodes - Fp1, Fp2, AF7, and AF8) in the prefrontal region Reference electrode: A2 Sampling rate: 512 Hz	Built-in real-time EEG signal enhancement to remove artifacts CCA used to decompose continuous signal into components GMM used to cluster features into groups, where outliers were removed	NA	*Power spectral density* (relative and absolute power) delta (1-3.5 Hz) theta (4-7.5 Hz) lower alpha (8-10 Hz) upper alpha (10.5-12 Hz)	Hochberg's sharpened Bonferroni adjusted significance values (primary significance level $p < 0.05$)
de la Salle, 2020	32 channel EasyCap system	Data was filtered (0.1-30 Hz), ocular-corrected, and inspected for artifacts	NA	*PFC cordance* (Average absolute and relative	

	Reference: average scalp reference sampling rate: 500 Hz	(voltages ± μV, faulty channels, drift) Minimum of 100 seconds of artifact-free data was required for participant inclusion		theta power from Fp1 and Fp2 electrodes at baseline and week 1) MRF cordance Average absolute and relative theta power from Fz, Fp2, F4, and F8 electrodes at baseline and week 1)	
Jaworska, 2019	32 channel EasyCap system Reference: average scalp reference sampling rate: 500 Hz	Ocular-corrected epochs excluded if voltage >±75 μV ln-transformation (normality) Min-Max scaling	Individuals with missing data (i.e., those without week 1 data) were removed (N = 2)	*eLORETA* *(Current source density measures from 84 Brodmann areas)* *Theta Cordance* (Average absolute and relative theta power from Fp1 and Fp2 electrodes at baseline and week 1)	KPCA
Wu, 2020	62-channel NeuroScan (EMBARC)	1) EEG data resampled to 250 Hz 2) 60-Hz AC line noise	Missing outcomes imputed using Bayesian regression	*SELSER* (each spatial filter transforms multichannel EEG data into a signal latent filter)	*SELSER* (performed under a sparse constraint on number of spatial filters, which serves to reduce

	256-channel Hydrocel Geodesic Sensor Net	artefact removed (CleanLine) 3) nonphysiological slow drifts removed using a 0.01-Hz high-pass filter 4) spectrally filtered EEG data re-referenced to the common average 5) bad epochs rejected by thresholding magnitude of each epoch 6) remaining artefacts removed using ICA		θ and α frequency bands	dimensionality)

First author, year	**Data utilized**	**Preprocessing / Quality Control Metrics**	**Imputation strategy**	**Feature extraction method** (*if applicable*)	**Feature selection method** (*if applicable*)
STUDIES USING NEUROIMAGING					

Braund, 2022	3T GE MRI Scanner Average time series extracted from 400 cortical regions and 36 regions from the subcortex. - BOLD time-series within each of these regions were correlated pair-wise with every other region and Fisher-Z transformed to create a 436x436 interregional functional correlation matrix for each participant.	Network-based statistic method used to analyse whole-brain network functional connectivity associated with neuroticism Covariates included age, sex, years of education, and baseline $HRSD_{17}$.	NA	Network-based statistics analysis identified a connectomic signature comprising 622 connections across 198 nodes in people with MDD, where greater neuroticism was associated with significantly higher functional connectivity (corrected p=0.10)	Filter-based approach was used, which is less prone to overfitting compared to wrapper methods.
Fan, 2020	3-Tesla structural (1x1x1 mm^3) and functional (3.2x3.2x3.1mm^3; TR = 2000 ms; TE=28 ms; 12 min) MRI	Slice timing correction, motion correction, intensity normalisation, brain masking, and registration of fMRI images to structural MRI and standard template.	NA	**Gray matter whole-brain parcellation using the A424 atlas** 1) NRS connectome is the pairwise connectivity of FNs affiliated modules at AA-24 and	2-tailed t-tests (FDR; $p < 0.05$)

		ICA-FIX used to identify and remove artefacts, followed by MGTR		AA-50 2) nodal strength (nS) is the average connectivity strength from one node to all other nodes in the brain 3) nodal internal NRS - average connectivity strength from one node to all other nodes within the same canonical connectivity FN 4) nodal external NRS - average connectivity strength from one node to all other nodes outside of its FN	
Fonzo, 2019	3-Tesla structural (1x1x1 mm^3) and functional (3.2x3.2x3.1mm^3; TR = 2000 ms; TE=28 ms; 12 min) MRI	FSL tools used to preprocess imaging data (FLIRT and FNIRT) Nuisance signals corresponding to segmented white matter and cerebrospinal fluid were regressed out of motion- corrected functional images.	MICE	Individual contrast maps specifying the difference in activation for iI–cI trials ROIs were mapped to seven previously identified functional networks according to the spatial overlap between each ROI and each network. (cortical, striatal, cerebellar, amygdala, anterior/posterior	Relevance Vector Machine (Bayesian evidence framework)

		A 6-mm full-width at half max isotropic smoothing kernel was then applied to preprocessed time series images to account for individual anatomical variability. Minimum behavioural accuracy during the emotional conflict task (≥ 80% of trials correct).		hippocampus, and thalamus)	
Klöbl, 2020	3T Siemens Biograph mMR system 71/77/100 min after infusion of study medication	1) correction of transient slice artefacts (ArtRepair) 2) slice-timing correction (SPM) 3) realignment (SPM) 4) reslicing of realigned images (SPM) 5) pre-smoothing with 4 mm FWHM Nuisance regression and frequency filtering were	For two patients, missing post-treatment scores were linearly interpolated from the visit before and after, and rounded up.	*Network-based statistical analysis* Differences in connectivity z-matrices between SSRI and placebo condition (significant threshold set to $p \leq 0.10$)	Median pearson correlation (significant threshold set to $p \leq 0.10$)

		performed for QC			
Koutsouleris 2017	3T structural MRI MNI-152 template smoothed with 8 mm Gaussian kernel	DARTEL (Automated tissue segmentation & high-dimensional stereotactic registration)	NA	PCA (PCs explaining study site with R^2 >.16 were removed)	PCA (20-25 PCs accounting for 80% of variance)
Nemati, 2020	*Sertraline* 3.0 T structural (1x1x1 mm3) and functional (3.2x3.2x3.1 mm3; TR=2000 ms; TE=28 min; 12 min at baseline and week 1 scans *Ketamine* 3.0 T structural (1x1x1 mm3) and functional (3x3x2.5 mm3; TR=3000 ms.; TE=30 ms.; 5min. immediately prior to infusion and 20 min.	FreeSurfer parcellation of structural scans, slice timing correction, motion correction, intensity normalisation, brain masking, and registration of fMRI images to structural MRI and standard template ICA-FIX used to identify and remove artefacts, followed by MGTR	NA	Full connectome computed as the pairwise Pearson correlation coefficients between these averaged time series, and subsequently transformed using a Fisher-z function. Nodal strength (nS) was computed as the average connectivity between each node and all other nodes within the full Connectome. Akiki-Abdallah cortical (AAc) atlas	Within- and between-network connectivity values (i.e., edges) were used in regression models (Pearson correlation) to identify NRS edges that positively or negatively predict the behavioural measure of interest ($p < 0.05$).

	during infusion starting at 20 min post administration)				
Nord, 2019	5 min T1-weighted anatomical scan (1 mm isotropic magnetization-prepared rapid gradient-echo) using a Siemens Avanto 1.5 Tesla MRI scanner with a 32-channel head coil. Echo time = 50 ms; repetition time per slice= 87 msec, in-plane resolution 2 x 2 mm. N-back working memory task performed during stimulation	For each time series, removed the first six volumes to allow for T1 equilibration, realigned the remaining volumes to the seventh volume, coregistered the volumes to each participant's anatomical scan, normalised into standardised space using the Montreal Neurological Institute template, and smoothed using an 8mm full width at half maximum Gaussian kernel.	Missing outcomes imputed using last observation carried forward	Intra-class correlation coefficients were calculated for each ROI that was significantly associated with clinical response Pre-randomization activation averaged within each ROI (the L-DLPFC for the n-back task; amygdalae and sgACC for the emotional faces task)	Constructed an independent samples t-test in SPM testing for the effect of group (active or sham) on each contrast and included percent change in HAM-D as a covariate in the model and the interaction (alpha = 0.05, two-tailed)
Sarpal, 2016	GE3-T scanner	FSL and AFNI	NA	ROIs were spherical regions	For every voxel located within

	5-min resting-state functional scans (150 whole-brain volumes) TR = 7.5 ms, TE = 3 ms, TI = 650 ms matrix = 256x256, FOV = 240 mm) producing 216 contiguous images (slice thickness = 1mm)	Rigid body motion correction performed with FLIRT Skull stripping performed with BET Images spatially smoothed with 6-mm FWHM Gaussian kernel High pass filter - 0.05 Hz Low pass filter - 0.1 Hz		with a radius of 3.5 mm around a seed voxel AFNI (3dfim+) used to create functional maps Mean time course of resting-state blood-oxygen-level-dependent activity was extracted from each seed region. he Fisher z-transformation was applied to the resulting correlation maps	gray matter (181,144 voxels total), the corresponding connectivity strength for each first-episode patient was entered into a univariate Cox regression analysis. Resulting z scores of this analysis for each voxel were placed in Montreal Neurological Institute standard brain space to create whole-brain maps Applied a threshold of $p<0.005$
Yip, 2019	Siemens Trio 3T scanner using a T2*-sensitive echo-planar image (EPI) gradient-echo pulse sequence (repetition time/echo time (TR/TE)=1500/27ms, flip angle=60º, field of view (FOV)=220x220mm, matrix=64x64, 3.4x3.4mm in-plane	Slice-time and motion correction performed using SPM8 All further analyses performed using BioImage Suite Several covariates of no interest were regressed from the data including linear and quadratic drifts, mean CSF signal, mean white-matter	NA	Whole-brain functional connectivity conducted using BioImage suite Network nodes defined using Shen 268-node brain atlas Mean time courses computed for each of the 268 nodes (i.e., average time course of voxels within the node) for use in node-by-node pairwise	Embedded feature selection (SVR)

	resolution, slice thickness=4mm with 1mm skip, 5mm effective slice thickness, 25 slices) During task performance, participants presented with one of six cues (win $1/$0/$5, lose $0/$1/$5) for 1000ms, indicating the amount of money to be won or lost on that trial, followed by a fixation cross (variable delay).	signal, and mean gray-matter signal Temporal smoothing using a Gaussian filter (approximate cutoff frequency=0.12 Hz) MID task runs were variance normalised and concatenated		Pearson's Correlation. r-values were transformed using Fisher's z-transformation to create symmetric 268x268 connectivity matrices in which each element of the matrix represents the strength of connection between two individual nodes.	

First author, year	**Data utilised**	**Preprocessing / Quality Control Metrics**	**Imputation strategy**	**Feature extraction method** (*if applicable*)	**Feature selection method** (*if applicable*)

STUDIES USING MULTIMODAL DATA					
Athreya, 2019	SNPs from blood samples identified during a GWAS for plasma kynurenine concentrations and 26 clinical and sociodemographic variables	Blood samples drawn at baseline; DNA genotyped using Illumina Human610-Quad Beadchips. Minor allele frequency, call rate and departure from Hardy-Weinberg equilibrium were evaluated.	NA	NA	NA
Ambrosen 2020	*MRI* Cohort A: T1-weighted sagittal MPRAGE images obtained with echo time (TE) 4 ms, repetition time (TR) 9.7 ms, flip angle 12°, field of view (FOV) 250 mm, matrix 256 × 256, 0.98 × 0.98 × 1 mm3 voxels, 170 slices. Cohort B: T1-weighted MPRAGE images were acquired	*MRI* Images processed using FreeSurfer Applied a 3T specific option for Talairach alignment *EEG* Processing performed using BESA software	Median imputation and probabilistic principal component analysis (PPCA) imputation	Segmentation of subcortical volumes involved neck removal, bias-field correction, brain extraction, tissue type segmentation, and FIRST CPTB consists of the prepulse inhibition (PPI), P50 suppression, mismatch negativity (MMN), and selective attention (SA) paradigms	NA

	with TE 3.93 ms, TR 1540 ms, flip angle 9°, FOV 256 mm, matrix 256 × 256, 1 mm isotropic voxels, 192 slices. Cohort C: T1-weighted FFE images were acquired with TE 4.6 ms, TR 10 ms, flip angle 8°, FOV 240 mm, matrix 304 x 299, acquired voxel size 0.79 x 0.80 x 0.80 mm3 and reconstructed voxel size 0.75 x 0.75 x 0.80 mm3, 200 slices. *EEG* 64 channel BioSemi continuous 70 dB played during PPI				
Crane, 2017	3.0 T GE Signa scanner using a	Data despiked using AFNI	NA	GIFT used to perform ICA	Beta weights from each event (Targets, Commissions,

	standard radio frequency coil and T2-weighted pulse sequence Repetition time = 2000 ms, echo time = 30 ms, flip angle = 90, field of view = 22 cm, 64 64 matrix, slice thickness = 4 mm, 29 slices.	Slice-time corrected in SPM8 and realigned in FSL using MCFLIRT Anatomical and functional images were co-registered and normalised to the T1-weighted structural image in MNI space using SPM8 Isotropic smoothing completed with a full-width at half-maximum filter of 5 mm3.		Pearson correlations were calculated between the behavioural results and the top five components for each behaviour for the individuals included in the ICA analysis	Rejections) were used as separate independent variables in a multiple regression.
Fonzo, 2018	Three behavioral paradigms that probe components of emotional reactivity and regulation, as well as a control task during an fMRI scan. 3-T GE Signa	During behavioural paradigms, measures of heart rate and respiration were collected and used to remove physiological noise from the time series. Global signal corresponding to segmented white matter	Analyses restricted to patients without missing data.	Whole brain exploratory analysis of conscious fear vs. neutral contrast. The a priori contrasts of interest were the differences in activation for conscious fear vs. neutral and for non-conscious (masked) fear vs.	The significance threshold was set at a family-wise error corrected $p < 0.05$ (two-tailed).

	scanner (T2*-weighted gradient echo (TR=2000 ms, TE=30 ms, flip angle = 80 degrees, field of view = 22 cm, 64x64 matrix). T1 structural scan used for anatomical localization of BOLD signal. Emotional Conflict Task: 148 presentations of an emotional face and instructed to identify underlying facial emotion (fearful or happy) while ignoring an overlying emotional distracting word. Reappraisal task: Presentation of 30 negative and 15 neutral photographs	and CSF was regressed out of motion-corrected functional images, which were isotropically smoothed with a 6 mm full-width half max (FWHM) to account for individual anatomical variability. Participants with a root mean square absolute movement > 3mm across the mean of the squared maximum displacements in each of the 6 estimated translational and rotational motion parameters for each functional run were excluded from further analysis for quality control purposes.		neutral, each allowing for the isolation of fear reactivity processes within a particular processing depth. Whole brain analyses were restricted to a probabilistic gray matter mask (> 40%) derived from an independent sample of healthy participants.	

	taken from the IAPS (International Affective Picture System) database. Gender Conflict Task: Task to identify face gender and ignore congruent or incongruent overlaid gender words.				
Joyce, 2021	Plasma metabolites in both the PGRN-AMPS and the CO-MED cohorts were measured by targeted metabolomics with the AbsoluteIDQ p180 assay Platform Six functionally validated pharmacogenomic SNP biomarkers in or near the	Metabolites were transformed by the Yeo-Johnson transformation then centred at zero and scaled to unit variance One of the six SNPs were genotyped in the CO-MED sample with a LooRsq > 99% and the remaining five SNPs were imputed using the Michigan Imputation Server with an imputation R2 > 97.5% and a call rate > 99%.	Features with ≥10% missingness and individuals missing ≥20% of features were excluded	NA	NA

	TSPAN5, ERICH3, DEFB1, and AHR genes, and related to MDD pathophysiology or citalopram/escitalopram response. PGRN-AMPS genotyping was done using Illumina human 610-Quad BeadChip CO-MED genotyping was done using Illumina Quad, Human Omni 2.5 bead chip Baseline clinical and sociodemographic variables				
Lee, 2018	First set of 13 SNPs (rs10803138,	NA	PAM algorithm	NA	NA

	rs11682175, rs6704641, rs6704768, rs215411, rs1106568, rs12522290, rs4129585, rs2514218, rs2239063, rs4702, rs12325245, and rs9636107) were from the 128 genome-wide significant associations for schizophrenia identified by the Schizophrenia Working Group of the Psychiatric Genomics Consortium The second set of 25 SNPs were from the top 25 results obtained from the GWAS for schizophrenia in the CATIE study although no SNP or combination of SNPs achieved genome-				

	wide statistical significance. 53 baseline clinical variables				
Nguyen, 2022	fMRI contrast maps parcellated into 200 functional brain regions during number-guessing trial, reward expectancy and prediction error 95 pretreatment clinical measures and demographic features acquired on the same day as imaging	fMRI data preprocessed using skull-stripping, head motion correction, spatial normalisation, and spatial smoothing with a 4-mm full width at half maximum kernel. Data augmentation, a process used in deep learning to reduce the likelihood of overfitting, was used, which generates additional image data by causing slight distortion to the original acquired images.	NA	Three contrast maps computed for each participant, quantifying brain activation in the initial anticipation phase of each number-guessing trial, reward expectancy (differential activation in rewarding vs. punishing trials), and prediction error (after wrong guesses). Each contrast map is parcellated into 200 functional brain regions using spatially constrained spectral clustering, yielding a total of 600 fMRI features for each participant.	Deep learning (feed-forward neural networks) were used, which incorporates embedded feature selection through stacking hidden layers.
Rajpurkar, 2020	Resting-state EEG (26-channels) was recorded for 2 minutes while participants were	NA	Patients with missing features (EEG or clinical) were excluded	A search was performed over various combinations of input features by altering the bands, time windows, and relative or absolute power of the EEG	Highest C index scores

	relaxed with eyes closed and eyes open from sites in 5 regions (frontal, temporal, central, parietal, and occipital) with a NuAmps system (Compumedics) and QuickCap (Compumedics).			features. Each feature is calculated at each of the 26 electrodes. Power of the EEG signals in each frequency range at each electrode site were extracted using the Welch method for spectral density estimation. Two additional features were computed: a frontal alpha asymmetry feature by subtracting alpha power for a left scalp site (F3) from the homologous right site (F4) and a beta- alpha ratio feature by taking the ratio of the beta features at each of the sites with the corresponding alpha features	
Rethorst, 2017	25 clinical variables, five baseline serum biomarkers (IL-1B, IL6, TNF-α, SHAPS, BDNF)	NA	NA	NA	Bootstrap estimated mean decrease in Gini

Taliaz, 2021	DNA samples were genotyped on arrays measuring 500,000 or more SNPs that tag most common variants in the human genome (43 features, 27 features are genetic variants that were segmented to 26 genetic components)	DNA was extracted from blood or lymphoblastoid cell lines and genotyped on arrays measuring 500,000 or more single-nucleotide polymorphisms (SNPs) that tag most common variants in the human genome. DNA samples were then genotyped using the Affymetrix© Human Mapping 500K Array and the Genome-Wide Human SNP Array 5.0	NA	Gene Ontology (GO) enrichment analysis	Embedded feature selection (SVM, Random Forest AdaBoost)

Supplementary Table S3 - Feature processing, selection, and extraction

1. Search Filter

Scopus

((*artificial* AND *intelligence*) OR (*supervised* AND *machine* AND *learning*)) **AND** ((*mental* AND *disorder*) OR (*mental* AND *disorders*) OR (*psychiatric* AND *disorder*) OR (*psychiatric* AND *disorders*) OR (*mental* AND *illness*) OR (*bipolar* AND *disorder*) OR (*schizophrenia*) OR (*depressive* AND *disorders*) OR (*anxiety* AND *disorders*) OR (*substance* AND *use* AND *disorder*) OR (*attention* AND *deficit* AND *disorder* AND *with* AND *hyperactivity*) OR (*personality* AND *disorders*) OR (*stress,* AND *disorders,* AND *post-traumatic*) OR (*trauma* AND *stressor* AND *related* AND *disorders*)) **AND** ((*clinical* AND *trials*) OR (*clinical* AND *trial*) OR (treatment AND selection) OR (treatment AND response) OR (treatment AND prediction)) **AND** (LIMIT-TO (DOCTYPE , *"ar"*)) AND (LIMIT-TO (SUBJAREA , *"MEDI"*) OR LIMIT-TO (SUBJAREA , *"COMP"*) OR LIMIT-TO (SUBJAREA , *"NEUR"*) OR LIMIT-TO (SUBJAREA , *"PSYC"*) OR LIMIT-TO (SUBJAREA , *"ENGI"*) OR LIMIT-TO (SUBJAREA , *"BIOC"*))

Results: 10900

Search Date: 2022-03-22

PubMed

((((((((((((("Artificial Intelligence"[Majr]) OR "Supervised Machine Learning"[Majr]) **AND** "Mental Disorders"[Majr]) OR "Anxiety Disorders"[Majr]) OR "Bipolar and Related Disorders"[Majr]) OR "Feeding and Eating Disorders"[Majr]) OR "Mood Disorders"[Majr]) OR "Personality Disorders"[Majr]) OR "Schizophrenia Spectrum and Other Psychotic Disorders"[Majr]) OR "Substance-Related Disorders"[Majr]) OR "Trauma and Stressor Related Disorders"[Majr]) **AND** "Clinical Trials as Topic"[Majr]) OR “Treatment response”[Other Term]) OR “treatment prediction”[Other Term]) OR “Treatment selection”[Other Term])

Results: 3471

Search Date: 2022-03-22

Web of Science

((((WC=(Supervised Machine Learning)) AND (WC=(Mental Disorders))) AND AB=(Treatment response)) OR AB=(treatment prediction)) OR AB=(Treatment selection)

Document Types: Articles

Web of Science Categories: Psychiatry or Neurosciences

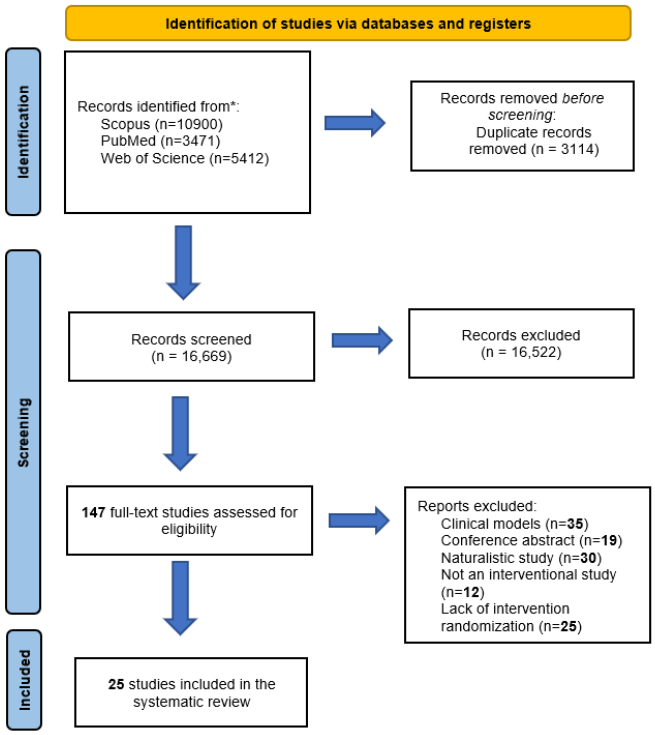
Results: 5412

Search Date: 2022-03-22

Total records (before duplicate removal): 19,723

Total records (duplicates removed):16,669

PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only



2. Quality assessment instrument development

We formed a group of multidisciplinary researchers from the fields of Neuroscience, Psychiatry, and Computer Science to develop a time efficient and practical assessment strategy to evaluate the quality of machine learning based healthcare research. For that purpose, we attempted to capture the reliability of the results presented in each study and identify practical ways that methodology may be improved.

This comprised nine methodological features, including sample representativeness, confounding variables, and outcome assessments, which were judged to be the most clinically pertinent components in machine learning-based healthcare research. Relevant considerations of each methodological feature are discussed in further detail in the next sections. The six remaining dimensions assess the quality and specific components of the machine learning approach that were used in a given study. In summary, this entails the algorithm or framework used, evidence that hyper-parameter optimization and feature selection procedures were used, whether authors provided details on how missing data and class imbalance problems were handled, the accuracy of a given model, and finally whether the model performance was tested in unseen data. These dimensions were qualitatively evaluated according to the information in section 3.

3. Quality assessment instrument domains

Methodological Feature	Considerations
1. Representativeness of the sample	Was the study representative of the heterogeneity observed in the target population? If not, was this related to the sampling method, insufficient sample size or inclusion/exclusion criteria?
2. Confounding variables	Did the study control for the most relevant confounding variables? If so, were covariates assessed using subjective or objective measures?
3. Outcome assessment	How were outcome measures assessed:

	A. Independent blind assessment (✓) B. Secure record (e.g., surgical records) (✓) C. Interview not blinded, self-report or medical record D. No description
4. Algorithm selection	Was the machine learning algorithm used to analyze the data clearly described and appropriate?
5. Feature selection	Did the study describe both feature selection and hyperparameter tuning? Which metrics were used?
6. Class imbalance	Did the authors address the class imbalance problem? Which method was utilized?
7. Missing data	Did the study describe how the authors handled missing data, including whether they were inputted or removed?
8. Performance/accuracy	Were the following performance metrics included for classification studies? A. Accuracy B. Sensitivity C. Specificity D. AUC E. PPV/NPV F. 95% Confidence intervals of performance metrics Or, alternatively, were one of the following performance metrics included for regression studies? A. Mean-squared error B. Mean-absolute error C. Root-mean-squared error
9. Testing/validation	Was the test dataset "unseen" in regard to model training? Was the model tested on a hold-out or an external dataset?

3.1. Representativeness of the sample

Machine learning models can deal with large amounts of data and the problem of heterogeneity. Therefore, there is less of a need to be restrictive with inclusion and exclusion criteria, relative to a traditional statistical approach examining significant effects at a group-level. Considering all studies included in the present review used data from randomized clinical trials, determined whether 1) performance was tested on an external sample with differences in inclusion/exclusion criteria, and 2) whether a training sample of ≥ 100 patients was used in model development.

3.2. Internal CV

To adequately control for confounding variables within machine learning models, it is important to ensure that these variables have a similar effect across the entire sample. To achieve this, randomization is an important step within the analysis. Often, the overall sample is randomly split into training and testing sets, and the analysis is repeated on the training dataset with different hyperparameters to maximize accuracy and minimize error. This is known as internal cross-validation. From here, if model performance is similar in the testing dataset, it presumes that potential confounding variables are uniformly distributed across the sample. Using this criteria, we evaluated whether the authors controlled for confounding variables.

3.3. Outcome assessment

How an outcome is defined has several important implications in a predictive model. Depending on the question or problem, a classification task may be appropriate, which uses a categorical outcome, or a regression task may be more relevant, where the outcome is continuous and numeric. A clinical instrument or questionnaire, for example, can be used as a numeric score or it can be transformed into a categorical outcome by using a cut-off score. We evaluated how authors assessed these outcomes, considering (A) independent blind assessments and secure records as high quality, (B) unblinded interview, self-report, or medical record as lower quality and (C) when no description was available.

3.4. Algorithm selection

There are several algorithms to choose from, with each relying on slightly different assumptions of the underlying data. Broadly speaking, there are linear (logistic regression, linear support vector machine), non-linear (Naive Bayes, K-Nearest Neighbors, Learning Vector Quantization)

tree-based (decision trees, random forest, xgboost) and neural network (convolutional neural

network, multilayer perceptrons) models, although others exist. Certain algorithms may be better suited to particular problems. For example, tree-based models such as random forest may be better suited to datasets with multicollinearity among features than linear-based models such as logistic regression. However, regularization parameters can be used in linear-based models (such as L2 regularization) to account for issues such as this.

Nevertheless, it is often difficult to determine beforehand which algorithms will lead to the highest model performance. Therefore, it is often a good strategy to compare the model performance of several algorithms. In this item, we evaluated whether the authors used an algorithm that is commonly used for the specific type of dataset, if several algorithms were compared, and if hyperparameter tuning was used.

The appropriateness of a machine learning algorithm was determined based on whether the specific data used in model development was congruent or incongruent with the strengths and limitations of the specific algorithm. For example, if a Gaussian process model was used, which is a non-sparse algorithm that loses efficiency in high dimensional spaces, in conjunction with a high-dimensional dataset, this algorithm would be deemed inappropriate for the input data. Conversely, Naive Bayes, which works well with high dimensional data would be considered an appropriate algorithm in such cases. Another example of an inappropriate model would be the use of convolutional neural networks for structural and tabular style datasets, as such algorithms are better suited to unstructured datasets. In cases where authors included both appropriate and inappropriate algorithms during model development, this consideration is scored with a “B”, alongside an asterisk to indicate which algorithms were inappropriate and why. Studies which only utilized one algorithm during model development that was deemed inappropriate received a score of “C”. Furthermore, studies are scored with a “B” if they did not compare multiple algorithms during model development and were scored as an “A” if they compared multiple algorithms that were deemed appropriate based on the candidate feature set.

3.5. Feature selection

A common problem in machine learning studies is the so-called small-n-large-p problem, also known as the curse of dimensionality, which occurs when there are more variables than examples in a dataset. Machine learning models created using these datasets are more prone to overfitting, which often results in overinflated performance in a training dataset, but much poorer performance in an external testing dataset. In addition, some algorithms cannot deal with more dimensions than examples. Highly correlated variables can also introduce more importance to a specific characteristic, decreasing the importance of the remaining variables. To circumvent these issues, a proper feature selection procedure, when applicable, should be done prior to

training or as part of the training procedure, such as it happens in embedded methods. The feature selection can be knowledge-driven or data-driven. In this item, we examined if the study used a proper feature selection (if applicable).

3.6. Class imbalance

Class imbalance occurs when the distribution of the outcome classes is highly unbalanced, i.e., when one outcome occurs much more frequently than the other outcome(s). This may result in a model with high accuracy but with very little clinical utility. For example, let us suppose that we have 95 occurrences of response in our dataset and only 5 occurrences of a nonresponse. Even if our model has 95% accuracy, it is useless if the model cannot detect the five instances of non-response high accuracy. In this item, we evaluated whether there was a class imbalance in the sample and if this problem was correctly addressed. This can be done using a series of methods, including (1) changing the metric of performance (accuracy, for example, is a poor form of evaluating imbalanced data sets; (2) resampling the data set by artificially increasing it (oversampling) or by removing examples from the majority class to create a more balanced data set (undersampling); (3) by generating more data with algorithms such as the Synthetic Minority Over-Sampling Technique (SMOTE); (4) by choosing algorithms that deal better with unbalanced classes, such as CART or random forests; (5) by using penalized models; or (6) by using anomaly and change detection. In cases where class imbalance was not relevant (balanced classes or regression models) this is scored as “yes”.

3.7. Missing data

It is critical to handle missing data since several algorithms cannot process incomplete data sets. Furthermore, it is also necessary to use an adequate imputation method to avoid introducing bias, which would otherwise lead to false conclusions if not addressed. It is important to report the amount of missing data in each variable, if these cases were excluded, or if the authors used an algorithm to input data and which algorithm/technique was used. Ideally, authors should provide a visual distribution of the patterns of missing data, such as aggregation plots, spinogram/spineplots, mosaic plots, etc. All these factors were evaluated in this section.

3.8. Performance/accuracy

Here, we evaluate whether the authors reported all relevant results and if they used the appropriate metrics. Studies informing only partial metrics may mask bias and flaws of the method, preventing the reader from fully understanding the relevance of the model. Confidence intervals should ideally be available for all performance metrics.

3.9. Testing/Validation

We can divide the machine learning process into three main components: training, validation, and testing. A training set allows the algorithm to learn and develop a predictive model. The validation set contains unseen data and is used to control for overfitting. Frequently, the same dataset is divided into training and validation sets. After a model is trained and validated, and shows consistent performance in both these steps, the model can be applied in an external and independent testing set. This allows us to see if the model can be generalized outside of the original sample. Some validation methods include holdout validation, k-fold, and leave one out cross validation.

A model that shows good performance in the training set but performs significantly poorer in the validation step is most likely due to overfitting - which occurs when the model relies more on the specific nuances and noise of the training dataset, resulting in poor accuracy in unseen data. In this item, we evaluated whether the authors properly tested and validated their models by taking steps to improve its generalizability. It is important to highlight that the use of cross-validation to evaluate performance should be discouraged when the data is large enough for a training-test split. Furthermore, the size of the test set should be sufficiently large for accuracy and other metrics to be estimated with high reliability.

4. Additional Methodological Considerations

4.1. Calculating a heterogeneity score in patients

A longstanding problem in clinical trials in psychiatry is patient heterogeneity. As such, while a novel medication may be highly effective for a subset of patients, it may fail placebo control in the presence of excessive treatment effect variability across the sample. Although we advocate for a shift towards machine-learning guided trials, calculating baseline patient heterogeneity scores at an individual level may foster more effective patient recruitment within traditional clinical trials in psychiatry. While this approach may exclude a subset of patients from larger trials, it would also provide greater flexibility in recruiting patients that have a higher likelihood of attaining treatment response. Furthermore, it is equally important to assess punitive mechanisms of insufficient treatment effects in patients who do not respond in initial feasibility trials.

4.2. Drug Discovery and Drug Repurposing

While drug discovery using ML applications remains largely unexplored in the context of mental health, a few preliminary studies suggest the feasibility of this. Zhao & So [1] describe a general approach to drug repurposing in psychiatry using the predictors of gene expression for profiles for each medication. L1 regularization is used to identify the top gene expression targets. Medications which are not approved for the disorder but have a high predicted probability of a good candidate, can be tested in prospective trials. Furthermore, Ekins et al. [2] detail an ML platform for end-to-end drug discovery and development. Issa et al. [3] also report a machine-learning guided modelling of biological processes in cancer to discover new disease-related targets, drug-phenotype associations, and discovery novel therapeutic targets. Additionally, Rodriguez et al. [4] showed that a machine learning framework applied to a list of genes can nominate drugs that may be repurposed for use in Alzheimer's disease. Considering that most medications in psychiatry, such as lithium, have been discovered by happenstance, ML models developed using high-quality biological data may help identify new therapeutic agents and repurpose currently approved medications for use in psychiatry.

4.3. Common Regularization Techniques

Regularization techniques, broadly speaking, are useful to decrease model complexity and improve model performance by applying various cost functions. Common regularization techniques applied to logistic regression models include L1 (ridge) and L2 (lasso) regularization. However, there are several available regularization methods, as described elsewhere [5].

4.3.1. L1 regularization

Briefly, Least Absolute Shrinkage and Selection Operator (LASSO) [6]adds a penalty term that is equivalent to the sum of the absolute value of coefficients. As such, LASSO shrinks less important feature coefficients to zero, which can work as a form of feature selection if we have more features than the number of individuals in our model [7]. The tuning parameter for LASSO is λ (alpha), and the lower the alpha, the more the model will resemble a standard linear regression model. A greater value of alpha places greater restrictions on the coefficients, leading to a sparser model [6].

4.3.2. L2 regularization

L2 regularization adds a penalty term that is equivalent to the squared magnitude of coefficients. By minimizing the sum of squares of coefficients, we can reduce the impact of correlated predictors. We can also introduce bias, which is referred to as lambda, so that our predictions are less sensitive to certain independent variables. When lambda corresponds to zero, the penalty is

also zero, and so we are essentially minimizing the sum of squared residuals. When lambda increases, our bias increases, however too much bias can also degrade model performance. [8]

Predicting adverse drug reactions in clinical trials

Recently it was shown that adverse drug reactions could be predicted with an AUC of 0.79-0.85 [9]. It was also shown that a deep learning model was able to learn molecular substructures that are specific to an adverse drug reaction [10]. In the context of psychiatry, adverse drug reactions are common, and this consideration also becomes salient when considering prospective clinical trials with novel compounds. Moreover, in another recent study, Yoo et al. [11] predicted sleep side effects from an 8-week, open-label trial of methylphenidate in pediatric ADHD with an accuracy of 86.1%. While preliminary, the key features in their model included fronto-striatal connectivity, and the SNPs DAT1, ADRA2A, and SLC6A2 [11].

In future studies, it is important to predict adverse reactions in interventional trials using cost-effective biological, clinical, and physiological data that can be applied to the clinic. Ideally, such models would be well-suited to small RCTs, such as a feasibility trial or Phase II study, to decrease the probability of adverse reactions in large-scale phase III trials. Similarly, patient dropouts remain a persistent issue in feasibility and pilot studies, which can render a trial underpowered. As such, there is a need for studies predicting medication tolerability, and whether a given patient is likely to drop-out prior to the end of treatment.

Predictive biomarker discovery

While there is a vast literature on predictive models and biomarkers in mental health, very few have been validated in clinical trials. Other fields of medicine, as described elsewhere, have used biological data, such as gene expression, to predict drug sensitivity [12]. This may be facilitated by the rapid evolution of low-cost, portable high-throughput single-cell RNA sequencing, which have been used for cell-specific biomarker discovery [13]. Importantly, new feature selection methods for biomarker discovery have recently been developed. For example, it was shown that a probabilistic generative model can reduce the high-dimensional space in single-cell gene expression data and provide uncertainty estimates [14]. However, to our knowledge, no predictive models have been conducted using next generation sequencing to inform punitive mechanisms and therapeutic targets in psychiatry. Among the available studies in mental health, Niculescu et al. found that a set of gene expression changes and clinical markers could predict suicidal ideation across psychiatric diagnoses with an AUC of 0.92 [15].

Supplementary Material References

1. Zhao, K. & So, H.-C. Drug Repositioning for Schizophrenia and Depression / Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. *IEEE J. Biomed. Heal. Informatics* **23**, 1304–1315 (2019).
2. Ekins, S. *et al.* Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **18**, 435–441 (2019).
3. Issa, N. T., Stathias, V., Dakshanamurthy, S., Surgery, C. & Comprehensive, L. Machine and Deep Learning Approaches for Cancer Drug Repuroposing. *Semin Cancer Biol.* **68**, 132–142 (2021).
4. Rodriguez, S. *et al.* Machine learning identifies candidates for drug repurposing in Alzheimer's disease. *Nat. Commun.* **12**, (2021).
5. Tian, Y. & Zhang, Y. A comprehensive survey on regularization strategies in machine learning. *Inf. Fusion* **80**, 146–166 (2022).
6. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *R. Stat. Soc.* **58**, 267–288 (1996).
7. Muthukrishnan, R. & Rohini, R. LASSO: A Feature Selection Technique In Predictive Modeling For Machine Learning. *IEEE Int. Conf. Adv. Comput. Appl.* (2016).
8. Buteneers, P., Caluwaerts, K., Dambre, J., Verstraeten, D. & Schrauwen, B. Optimized Parameter Search for Large Datasets of the Regularization Parameter and Feature Selection for Ridge Regression. *Neural Process Lett* 403–416 (2013) doi:10.1007/s11063-013-9279-8.
9. Valeanu, A., Damian, C., Marineci, C. D. & Negres, S. The development of a scoring and ranking strategy for a patient- tailored adverse drug reaction prediction in polypharmacy. *Scientifi* **10**, 1–11 (2020).
10. Dey, S., Luo, H., Fokoue, A., Hu, J. & Zhang, P. Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinformatics* **19**, 1–13 (2018).
11. Yoo, J. H. *et al.* Prediction of sleep side effects following methylphenidate treatment in ADHD youth. *NeuroImage Clin.* 102030 (2019) doi: 10.1016/j.nicl.2019.102030.
12. Vamathevan, J., Clark, D., Czodrowski, P. & Cleveland, L. S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).

13. Gierahn, T. M. *et al*. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).

14. Ding, J., Condon, A. & Shah, S. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, (2018).

15. Niculescu, A. B. *et al.* Understanding and predicting suicidality using a combined genomic and clinical risk assessment approach. *Mol. Psychiatry* 1–20 (2015) doi:10.1038/mp.2015.112.

References

1. McCormack, L. *et al.* Communication and dissemination strategies to facilitate the use of health-related evidence. *Evidence report/technology assessment* (2013) doi:10.23970/ahrqepcerta213.
2. Devereaux, P. J. & Yusuf, S. The evolution of the randomized controlled trial and its role in evidence-based decision making. *J. Intern. Med.* 105–113 (2003).
3. Moher, D. & Olkin, I. Meta-analysis of Randomized Controlled Trials A Concern for Standards. *JAMA* **274**, 1962–1964 (1992).
4. Beckmann, J. S. & Lew, D. Reconciling evidence-based medicine and precision medicine in the era of big data: Challenges and opportunities. *Genome Medicine* (2016) doi:10.1186/s13073-016-0388-7.
5. Cipriani, A. *et al.* Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet* **391**, 1357–1366 (2018).
6. Leucht, S. *et al.* Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: A multiple-treatments meta-analysis. *The Lancet* **382**, 951–962 (2013).
7. Howes, O. D., Thase, M. E. & Pillinger, T. Treatment resistance in psychiatry: state of the art and new directions. *Molecular Psychiatry* (2021) doi:10.1038/s41380-021-01200-3.
8. López-Muñoz, F. & Alamo, C. *Monoaminergic Neurotransmission: The History of the Discovery of Antidepressants from 1950s Until Today*. *Current Pharmaceutical Design* vol. 15 (2009).
9. Harmer, C. J., Duman, R. S. & Cowen, P. J. How do antidepressants work? New perspectives for refining future treatment approaches. *The Lancet Psychiatry* vol. 4 409–418 (2017).
10. Fan, J., Han, F. & Liu, H. Challenges of Big Data analysis. *National Science Review* (2014) doi:10.1093/nsr/nwt032.
11. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* (2015) doi:10.1038/nature14541.
12. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems* (2014) doi:10.1186/2047-2501-2-3.
13. Liberati, A. *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. in *Journal of clinical epidemiology* (2009). doi:10.1016/j.jclinepi.2009.06.006.

14. Fonzo, G. A. *et al.* PTSD psychotherapy outcome predicted by brain activation during emotional reactivity and regulation. *American Journal of Psychiatry* **174**, 1163–1174 (2017).

15. Athreya, A. P. *et al.* Pharmacogenomics-Driven Prediction of Antidepressant Treatment Outcomes: A Machine-Learning Approach With Multi-trial Replication. *Clinical Pharmacology and Therapeutics* **106**, 855–865 (2019).

16. de la Salle, S., Jaworska, N., Blier, P., Smith, D. & Knott, V. Using prefrontal and midline right frontal EEG-derived theta cordance and depressive symptoms to predict the differential response or remission to antidepressant treatment in major depressive disorder. *Psychiatry Research: Neuroimaging* **302**, 111109 (2020).

17. Amminger, G. P. *et al.* Predictors of treatment response in young people at ultra-high risk for psychosis who received long-chain omega-3 fatty acids. *Translational Psychiatry* **5**, 3–9 (2015).

18. Hou, J. *et al.* Subgroup Identification in Personalized Treatment of Alcohol Dependence. *Alcoholism: Clinical and Experimental Research* **39**, 1253–1259 (2015).

19. Maciukiewicz, M. *et al.* GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of Psychiatric Research* **99**, 62–68 (2018).

20. Al-Kaysi, A. M. *et al.* Predicting tDCS treatment outcomes of patients with major depressive disorder using automated EEG classification. *Journal of Affective Disorders* **208**, 597–603 (2017).

21. Cao, Z. *et al.* Identifying Ketamine Responses in Treatment-Resistant Depression Using a Wearable Forehead EEG. *IEEE Transactions on Biomedical Engineering* **66**, 1668–1679 (2019).

22. Jaworska, N., de La Salle, S., Ibrahim, M. H., Blier, P. & Knott, V. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. *Frontiers in Psychiatry* (2019) doi:10.3389/fpsyt.2018.00768.

23. Wu, W. *et al.* An electroencephalographic signature predicts antidepressant response in major depression. *Nature Biotechnology* **38**, 439–447 (2020).

24. Braund, T. A. *et al.* Intrinsic Functional Connectomes Characterize Neuroticism in Major Depressive Disorder and Predict Antidepressant Treatment Outcomes. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **7**, 276–284 (2022).

25. Fan, S. *et al.* Pretreatment Brain Connectome Fingerprint Predicts Treatment Response in Major Depressive Disorder. *Chronic Stress* **4**, (2020).

26. Fonzo, G. A. *et al.* Brain regulation of emotional conflict predicts antidepressant treatment response for depression. *Nature Human Behaviour* **3**, 1319–1331 (2019).

27. Klöbl, M. *et al.* Predicting Antidepressant Citalopram Treatment Response via Changes in Brain Functional Connectivity After Acute Intravenous Challenge. *Frontiers in Computational Neuroscience* **14**, (2020).

28. Koutsouleris, N. *et al.* Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry* **3**, 935–946 (2016).

29. Nemati, S. *et al.* A Unique Brain Connectome Fingerprint Predates and Predicts Response to Antidepressants. *iScience* **23**, 100800 (2020).

30. Nord, C. L. *et al.* Neural predictors of treatment response to brain stimulation and psychological therapy in depression: a double-blind randomized controlled trial. *Neuropsychopharmacology* **44**, 1613–1622 (2019).

31. Sarpal, D. K. *et al.* Baseline striatal functional connectivity as a predictor of response to antipsychotic drug treatment. *American Journal of Psychiatry* **173**, 69–77 (2016).

32. Yip, S. W., Scheinost, D., Potenza, M. N. & Carroll, K. M. Connectome-based prediction of cocaine abstinence. *American Journal of Psychiatry* **176**, 156–164 (2019).

33. Ambrosen, K. S. *et al.* A machine-learning framework for robust and reliable prediction of short- and long-term treatment response in initially antipsychotic-naïve schizophrenia patients based on multimodal neuropsychiatric data. *Translational Psychiatry* **10**, (2020).

34. Crane, N. A. *et al.* Multidimensional prediction of treatment response to antidepressants with cognitive control and functional MRI. *Brain* **140**, 472–486 (2017).

35. Lee, B. S. *et al.* A computational algorithm for personalized medicine in schizophrenia. *Schizophrenia Research* **192**, 131–136 (2018).

36. Joyce, J. B. *et al.* Multi-omics driven predictions of response to acute phase combination antidepressant therapy: a machine learning approach with cross-trial replication. *Translational Psychiatry* **11**, (2021).

37. Nguyen, K. P. *et al.* Patterns of Pretreatment Reward Task Brain Activation Predict Individual Antidepressant Response: Key Results From the EMBARC Randomized Clinical Trial. *Biological Psychiatry* **91**, 550–560 (2022).

38. Rajpurkar, P. *et al.* Evaluation of a Machine Learning Model Based on Pretreatment Symptoms and Electroencephalographic Features to Predict Outcomes of Antidepressant Treatment in Adults With Depression: A Prespecified Secondary Analysis of a Randomized Clinical Trial. *JAMA Netw Open* **3**, e206653 (2020).

39. Rethorst, C. D., South, C. C., Rush, A. J., Greer, T. L. & Trivedi, M. H. Prediction of treatment outcomes to exercise in patients with nonremitted major depressive disorder. *Depression and Anxiety* **34**, 1116–1122 (2017).

40. Taliaz, D. *et al.* Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data. *Translational Psychiatry* **11**, (2021).

41. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. in *Proceedings - International Conference on Pattern Recognition* 3121–3124 (2010). doi:10.1109/ICPR.2010.764.

42. Wu, W. *et al.* An electroencephalographic signature predicts antidepressant response in major depression. *Nature Biotechnology* **38**, 439–447 (2020).

43. Trivedi, M. H. *et al.* Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. *Journal of Psychiatric Research* vol. 78 11–23 (2016).

44. Roelofs, R. *et al. A Meta-Analysis of Overfitting in Machine Learning*. https://www.kaggle.com/kaggle/meta-kaggle.

45. Bernau, C. *et al.* Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30**, (2014).

46. Dobbin, K. K. & Simon, R. M. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics* **4**, (2011).

47. Bruder, G. E. *et al.* Electroencephalographic Alpha Measures Predict Therapeutic Response to a Selective Serotonin Reuptake Inhibitor Antidepressant: Pre- and Post-Treatment Findings. *Biological Psychiatry* **63**, 1171–1177 (2008).

48. 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. *Journal of Affective Disorders* **256**, 132–142 (2019).

49. de la Salle, S., Jaworska, N., Blier, P., Smith, D. & Knott, V. Using prefrontal and midline right frontal EEG-derived theta cordance and depressive symptoms to predict the differential response or remission to antidepressant treatment in major depressive disorder. *Psychiatry Research - Neuroimaging* **302**, (2020).

50. Arns, M., Drinkenburg, W. H., Fitzgerald, P. B. & Kenemans, J. L. Neurophysiological predictors of non-response to rTMS in depression. *Brain Stimulation* **5**, 569–576 (2012).

51. Zandvakili, A. *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. *Journal of Affective Disorders* **252**, 47–54 (2019).

52. Hahn, T. *et al.* Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. *JAMA Psychiatry* **72**, 68–74 (2015).

53. Cao, B. *et al.* Predicting individual responses to the electroconvulsive therapy with hippocampal subfield volumes in major depression disorder. *Scientific Reports* **8**, 1–8 (2018).

54. Leaver, A. M. *et al.* Fronto-temporal connectivity predicts ECT outcome in major depression. *Frontiers in Psychiatry* **9**, 1–11 (2018).

55. Cash, R. F. H. *et al.* A multivariate neuroimaging biomarker of individual outcome to transcranial magnetic stimulation in depression. *Human Brain Mapping* **40**, 4618–4629 (2019).

56. Gong, J. *et al.* Predicting response to electroconvulsive therapy combined with antipsychotics in schizophrenia using multi-parametric magnetic resonance imaging. *Schizophrenia Research* **216**, 262–271 (2020).

57. Ge, R., Downar, J., Blumberger, D. M., Daskalakis, Z. J. & Vila-Rodriguez, F. Functional connectivity of the anterior cingulate cortex predicts treatment outcome for rTMS in treatment-resistant depression at 3-month follow-up. *Brain Stimulation* **13**, 206–214 (2020).

58. Xi, Y. bin *et al.* Neuroanatomical Features That Predict Response to Electroconvulsive Therapy Combined With Antipsychotics in Schizophrenia: A Magnetic Resonance Imaging Study Using Radiomics Strategy. *Frontiers in Psychiatry* **11**, 1–9 (2020).

59. Takamiya, A. *et al.* Predicting Individual Remission after Electroconvulsive Therapy Based on Structural Magnetic Resonance Imaging: A Machine Learning Approach. *Journal of ECT* **36**, 205–210 (2020).

60. Maciukiewicz, M. *et al.* GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of Psychiatric Research* **99**, 62–68 (2018).

61. Friedman, J. H. & Roosen, C. B. An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research* **4**, 197–217 (1995).

62. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* (2020) doi:10.1038/s42256-019-0138-9.

63. Wu, Y. Ultrahigh Dimensional Feature Selection: Beyond The Linear Model. *Journal of Machine Learning Research* **10**, 2013–2038 (2009).

64. Liu, W. & Li, R. Variable Selection and Feature Screening. in *Macroeconomic Forecasting in the Era of Big Data Theory and Practice* (eds. Liu, W. & Li, R.) 293–326 (2019).

65. Zhu, L. P., Li, L., Li, R. & Zhu, L. X. Model-free feature screening for ultrahigh-dimensional data. *J Am Stat Assoc* **106**, 1464–1475 (2011).

66. Li, X., Tang, N., Xie, J. & Yan, X. A nonparametric feature screening method for ultrahigh-dimensional missing response. *Computational Statistics and Data Analysis* **142**, (2020).

67. Guo, X., Ren, H., Zou, C. & Li, R. Threshold Selection in Feature Screening for Error Rate Control. *J Am Stat Assoc* (2022) doi:10.1080/01621459.2021.2011735.

68. Khalid, S. *A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning*. vol. 372 www.conference.thesai.org (2014).

69. Bonidia, R. P. *et al.* Feature extraction approaches for biological sequences: A comparative study of mathematical features. *Briefings in Bioinformatics* vol. 22 (2021).

70. Barandas, M. *et al.* TSFEL: Time Series Feature Extraction Library. *SoftwareX* **11**, (2020).

71. Naimi, A. I. & Balzer, L. B. Stacked generalization: an introduction to super learning. *European Journal of Epidemiology* **33**, 459–464 (2018).

72. Sesmero, M. P., Ledezma, A. I. & Sanchis, A. Generating ensembles of heterogeneous classifiers using Stacked Generalization. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **5**, 21–34 (2015).

73. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* vol. 180 68–77 (2018).

74. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning with Applications in R Second Edition*. (2021).

75. Braund, T. A. *et al.* Intrinsic Functional Connectomes Characterize Neuroticism in Major Depressive Disorder and Predict Antidepressant Treatment Outcomes. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **7**, 276–284 (2022).

76. Wagner, E. *et al.* Efficacy and safety of clozapine in psychotic disorders—a systematic quantitative meta-review. *Translational Psychiatry* vol. 11 (2021).

77. Farooq, S. & Taylor, M. Clozapine: Dangerous orphan or neglected friend? *British Journal of Psychiatry* vol. 198 247–249 (2011).

78. Nøhr, A. K. *et al.* A large-scale genome-wide gene expression analysis in peripheral blood identifies very few differentially expressed genes related to antidepressant treatment and response in patients with major depressive disorder. *Neuropsychopharmacology* **46**, 1324–1332 (2021).

79. Gierahn, T. M. *et al.* Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods* **14**, 395–398 (2017).

80. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications* **9**, (2018).

81. Hill, A. T., Rogasch, N. C., Fitzgerald, P. B. & Hoy, K. E. TMS-EEG: A window into the neurophysiological effects of transcranial electrical stimulation in non-motor brain regions. *Neuroscience and Biobehavioral Reviews* vol. 64 175–184 (2016).

82. Voineskos, D. *et al.* Altered Transcranial Magnetic Stimulation–Electroencephalographic Markers of Inhibition and Excitation in the Dorsolateral Prefrontal Cortex in Major Depressive Disorder. *Biological Psychiatry* **85**, 477–486 (2019).

83. Ferrari, M. & Quaresima, V. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage* vol. 63 921–935 (2012).

84. Chen, W. L. *et al.* Functional Near-Infrared Spectroscopy and Its Clinical Application in the Field of Neuroscience: Advances and Future Directions. *Frontiers in Neuroscience* vol. 14 (2020).

85. Husain, S. F. *et al.* Validating a functional near-infrared spectroscopy diagnostic paradigm for Major Depressive Disorder. *Scientific Reports* **10**, (2020).

86. Fagherazzi, G., Fischer, A., Ismael, M. & Despotovic, V. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers* vol. 5 78–88 (2021).

87. Mundt, J. C., Vogel, A. P., Feltner, D. E. & Lenderking, W. R. Vocal acoustic biomarkers of depression severity and treatment response. *Biological Psychiatry* **72**, 580–587 (2012).

88. Passos, I. C. & Mwangi, B. Machine learning-guided intervention trials to predict treatment response at an individual patient level: an important second step following randomized clinical trials. *Molecular Psychiatry* (2018) doi:10.1038/s41380-018-0250-y.

Chapter 9: Discussion

9.1. Summary of Findings

In the context of this thesis, our results showed that: 1) evidence-based risk factors, protective factors, and treatment status variables were able to prognosticate prospective physical aggression at an individual level, and that there was a statistically significant difference in error rate between a model comprising these variables and clinical judgement, relative to clinical judgement alone; 2) prognostic models of clinical and violent outcomes in psychiatry have largely focused on clinical and sociodemographic variables, show similar performance between identifying true positives and true negatives, although the error rate of models are still high, suggesting that further refinement is needed prior to implementing such models clinically; 3) within treatment response prediction models in MDD using EEG, greater performance was observed in predicting response to rTMS, relative to antidepressants, and across models, greater sensitivity (true positives), were observed relative to specificity (true negatives), suggesting that EEG prediction models thus far are better able to identify non-responders than responders; and 4) across randomized clinical trials using data-driven biomarkers in predictive models, based on the consistency of performance across models with large sample sizes, the highest degree of evidence was in predicting response to sertraline and citalopram using fMRI features. Importantly, a subset of these models has been replicated in independent datasets, largely maintaining meaningful but modest predictive accuracy, which suggests the potential for their scalability as classification tools. We also highlight that machine-learning guided intervention trials are lacking in psychiatry and propose a methodological pipeline to conduct prospective

machine-learning guided trials according to best practices and provide strategies to improve the interpretability and generalizability of predictive models.

9.2. Significance and General Discussion

In chapter 2, the predictive HARM models were developed using empirically supported risk factors of violence as candidate features, in conjunction with demographic variables, protective factors, and variables related to the course of treatment. These models were compared against clinician rated CLV and combined models, to evaluate their comparative performance, as briefly discussed in Chapter 6.1. Of note, clinician rated CLV, and HARM models largely showed contrary predictive values, where CLV models were better at identifying true positives, and HARM models were better at identifying true negatives, respectively. Moreover, although no statistically significant differences in error rates were observed between HARM and CLV models (McNemar's chi-square (χ^2) = 2.37, p=0.123), a statistically significant difference was found between CLV and combined models (McNemar's χ^2= 10.22, p= 0.001). As such, data-driven HARM variables, in conjunction with clinical judgement, appear to show better performance in discriminating physical aggression from non-aggression within patients with schizophrenia in forensic settings, better than clinical judgement alone. Moreover, CLV alone showed poor sensitivity across models (57.14-61.50%), indicating a high degree of false positives. Altogether, these results suggest that data-driven HARM models may show utility as an adjunct to clinical judgement, since these models are better able to identify true negative instances of non-aggression among patients, while clinical judgement is better able to identify true positive instances, relative to the HARM models. Moreover, several potential protective

factors emerged, including engagement in treatment programs, positive attitude, social support, family support, and medication adherence.

In contrast with existing actuarial tools, such as the VRAG [1] and HCR-20 [2], which consider a linear additive combination of variables to assess individual prospective risk, the HARM models incorporate a data-driven approach that allow for a non-linear weighting of importance between features, while also relying on theoretically sound and evidence-based risk factors, protective factors, and variables related to course of treatment. Moreover, the HARM models showed improvements in AUC relative to existing risk assessment tools, in predicting physical aggression at 4-month (AUC: 0.669-0.928), 12-month (AUC: 0.701-0.913), and 18-month (AUC:0.597-0.870) follow-up. Additionally, the HARM models incorporate additional performance measures, including sensitivity, specificity, balanced accuracy, overall model accuracy, as well as PPV and NPV, to better elucidate the goodness of fit of the models. While prospective validation is required, and a relatively small sample size was used in model development, machine learning models may show utility to improve the accuracy of risk prediction for individualised care of patients with schizophrenia in forensic settings.

In chapter 3, within a meta-analysis and systematic review, prognostic models of criminal and violent outcomes in psychiatry were assessed. Across eighteen models, accuracy ranged from 67.83-82%, with important variables in criminal outcome models including age at first crime, substance use disorder, cluster B personality disorder, prior criminality, a high number of stressors, and childhood trauma. Furthermore, models predicting violent behaviour were more variable, ranging from 58.25-92.1%, with important clinical features including confusion, irritability, threats, recently attacking objects, child abuse, physical neglect, and callous affect.

Overall, studies thus far have largely focused on electronic health record data, such as sociodemographic, clinical, and treatment-related variables. Within a bivariate meta-analysis of diagnostic accuracy, comprising this category of input features, the AUC for predicting violent and criminal outcomes in psychiatry was 0.816 (95% Confidence Interval (CI): 70.57-88.15), with a partial AUC of 0.773, average sensitivity of 73.33% (95% CI: 64.09-79.63), and average specificity of 72.90% (95% CI: 63.98-79.66), respectively. As such, based on available evidence, although clinical and sociodemographic variables appear to show discriminative capabilities above chance, they show a sizeable false positive (average: 26.67, 95% CI: 11.85-35.91), and false negative rate (average: 27.10, 95% CI: 20.34-36.02). Therefore, new approaches are warranted. For example, clinical and sociodemographic models may benefit from incorporating additional numeric clinical scales, that could perhaps be more conducive to capturing changes over time, relative to categorical or binary features. Additionally, it may be useful to incorporate features with greater granularity, such as a time series analyses with actigraphy and other wearable sensors, to potentially identify more time sensitive features, that may be more conducive to clinical intervention. Additionally, there are a lack of studies incorporating biological features to predict criminal and violent outcomes in psychiatry, as efforts thus far have only involved a set of single nucleotide polymorphisms and resting-state regional cerebral blood flow. Other modalities, such as EEG, task-based fMRI, or blood-based biomarkers, may also prove to be useful, although prospective studies are needed to elucidate this.

In chapter 4, we briefly consider the long-term ramifications of individualized prognostic models of clinical and violent outcomes within psychiatric disorders, which should be weighed against the strengths and weaknesses of available tools that are currently clinically implemented. To

further clarify, it is important to ensure that patient privacy is respected when scaling such models to clinical settings, for instance, how is training data stored, and shared, and whether data is stored centrally or cloud based, both of which may present a risk of personal data breaches involving highly sensitive criminal and healthcare records [3]. However, recent methods have been developed, including swarm learning, a decentralized machine learning approach that integrates blockchain-based peer-to-peer networking, edge computing, and maintains confidentiality without the need for a centralized hub [4]. Furthermore, another pertinent consideration is how these models will be implemented clinically. Namely, how will the care of patients who are predicted to be high risk of criminal and violent outcomes change? If these tools are used to triage patients who are predicted to be violent in the immediate future as requiring more cautious care, and reallocating resources including security and additional staff from patients who are not predicted to be an immediate or short-term danger, than this may potentially be feasible. However, in cases where predictive models are used to determine an individual's eligibility for privileges to enter into the community, it becomes more challenging ethically to implement, particularly if the false positive rate is high, as was the case within the HARM models described in Chapter 2. Altogether, while we advocate for more precise personalized prognostic models of criminal and violent outcomes in patients in order to improve targeted prevention, and overall clinical outcomes, we believe that precision psychiatry pertaining to the intersection of criminality, violence, and psychiatric patients who disproportionately have been diagnosed with schizophrenia [5], requires a precision ethics approach.

Similarly, in chapter 5, we open the discussion on predictive models of treatment response in psychiatry, with the use case of ketamine, a rapid acting antidepressant [6]. In terms of peripheral

markers, only a handful of studies have investigated peripheral blood biomarkers. Within a recent systematic review and meta-analysis of blood-based biomarkers of antidepressant response to ketamine and esketamine, no consistent associations were found between baseline levels of blood biomarkers and response to ketamine. Among longitudinal analyses, the only consistent finding was that ketamine responders showed significant increases in brain-derived neurotrophic factor (BDNF), relative to pre-treatment levels SMD [95% CI] = 0.26 [0.03, 0.48], p = 0.02), while non-responders did not show significant changes in BDNF (SMD [95% CI] = 0.05 [−0.19, 0.28], p = 0.70) [7].

Since ketamine and esketamine appear to show a divergent mechanism of action from traditional antidepressants [8], it is argued that there remains an unmet for need prospective clinical trials to investigate biological predictors of treatment response, and to examine the precision and true negatives of such biological predictors within classification models. Moreover, considering discrepancies across studies in the exact threshold of treatment response to discriminate responders and non-responders, as discussed further in chapter 6, regression models used to predict change scores may also be warranted. In chapter 6, while there is a great deal of promise in using EEG within machine learning models to predict treatment response in MDD, there does not appear to be a consensus on collection methods, or consistent physiological markers of response to antidepressants, or rTMS across studies. Given the complexity of MDD, and the likelihood of heterogeneity in important features across patients, the field may require a conceptual shift away from the search for singular biomarkers, towards the use of composite features, identified using multivariate models. As such, it may be the case that no singular neurophysiological biomarker will demonstrate the sensitivity and specificity required to guide

treatment selection in MDD. Rather, a composite biomarker comprising a series of distinct, but mutually informative features, may serve to both improve our mechanistic understanding of treatment response, and appropriately model this phenomenon. However, it is important to highlight that multimodal feature combinations carry several additional considerations. Namely, if complex approaches such as source localization are required to provide meaningful accuracy, this may provide a significant challenge in the clinical implementation of such models. Additionally, while resting-state features provide greater scalability relative to EEG activation patterns during specific tasks, the latter may inform features that could perhaps be more sensitive and specific in modelling clinical improvement in response to a given treatment.

In chapter 7, all studies predicting treatment response in randomized clinical trials using previously collected data, which necessitates a caution of their clinical implementation without adequate prospective validation. While RCTs have provided important insights into group-level statistics, they fail to yield individualized findings or account for patient heterogeneity. As such, we advocate for a new trial design to occur following the successful completion of an RCT. We refer to this as a *machine-learning precision trial.* Using standard RCT data within machine learning models garner two major limitations: (1) The sample included in the RCTs are not fully representative of the real clinical population with a specific disorder and (2) a considerable amount of the sample size is dedicated to a placebo condition, which may be better allocated towards an active arm from a modelling perspective.

Machine-learning precision trials must therefore possess three distinct components from traditional RCTs: (1) The vast majority of participants (≤90%) are allocated to the active treatment, and a small subset of patients (≥10%) are allocated to a placebo or sham control. This

allows for testing the specificity of biomarkers identified within the treatment arm; (2) greater flexibility in inclusion and exclusion criteria to increase the external validity of the trial, and reflect heterogeneous patients seen in the clinic, and (3) randomizing patients to medication dosages in the therapeutic range known to be effective, so that machine learning models can be trained to determine more individualized dosages based on patient characteristics.

With respect to the second consideration, it is important to note that while patient idiosyncrasies are commonly observed in real-world clinical settings, such as comorbidities, are common exclusion criteria in RCTs, greater flexibility in exclusion criteria may help to provide a more realistic appraisal of the generalizability and clinical utility of machine-learning precision trials.

Furthermore, although decreasing the sample size of individuals allocated to placebo conditions is required to maximize the sample in the active arm, it may be useful to retain a small proportion of the sample (approximately 10-20%), to be given an inert substance or sham condition, to determine the specificity of features relative to placebo. Additionally, other methods can be useful to control for placebo related features, such as utilising principal component analysis (PCA) [9] to identify the components explaining the majority (≥90%) of variance in predicting response to placebo and using a method such as multivariate adaptive regression splines (MARS) [10], where placebo related variance is imputed in the forward pass and removed from the set of candidate features in the backwards pass.

9.3. Limitations

Within chapter 2, although the study benefits from a longitudinal design, and showed similar variable importance across timepoints, a low base-rate of aggressive incidents was observed at 4-

, 12-, and 18-month follow-up. As such, future studies with larger sample sizes will be required to determine the replicability of predicting longitudinal physical aggression in patients with schizophrenia in forensic settings. Considering that the study used binary classification tasks, alongside baseline variables, to predict physical aggression, no hypothesis testing was performed, and as such, statistical power cannot be calculated. Since the present study had a low base rate of physical aggression, and relatively small sample size, it is possible that model accuracy is inflated.

Additionally, it is important to consider that these models were developed in a specific at-risk cohort of patients with schizophrenia who have a history of criminal offences. As such, these models may not be generalizable to detect aggressive behaviours in schizophrenia in general. Moreover, our models were developed largely using categorical features, which were transformed into binary variables using one-hot encoding [11]. While several models were used that can handle multicollinearity, other methods, such as transforming features into principal components [44], can be used to derive a set of uncorrelated variables.

With respect to chapter 3, Currently, the field of predicting crime and violent related outcomes using machine learning techniques remain in its infancy. As such, there is a lack of studies validating model performance using independent cohorts. Furthermore, it is important to note that model accuracy should be considered alongside several other factors, such as the input features used, the preprocessing pipeline, feature selection method, model optimization strategy, and the validation procedure. Furthermore, data-driven approaches to feature selection can be useful in many cases, since it does not require knowledge derived from pre-existing literature to

manually select important variables [12–14]. Of note, the absence of a formalized feature selection strategy was observed across a subset of studies.

Additionally, only two studies developed separate models to assess potential differences in performance between men and women using the same variables. Rossellini et al. reported an AUC of 0.74 for men and an AUC of 0.82 for women in predicting violent crime [15]. Additionally, the same authors also investigated predictors of major violent crime and reported an AUC of 0.81 for both models in men, and an AUC of 0.80-0.82 for both models in women. Based on these studies, it is still unclear whether biological sex or gender play a key role in deciding which features should be included within a predictive machine learning model.

In reference to chapter 6, there is a need for greater emphasis on testing model performance with independent samples, greater consistency in sample collection and model development, and an increased focus on replicating features identified in previous models. Additionally, nine studies (60%) included in the present meta-analysis and systematic review did not test accuracy in holdout data, relying instead on internal cross-validation, which may lead to overoptimistic performance metrics. Furthermore, most studies (57.1%) utilised data from open-label trials lacking adequate double-blind procedures, and as such, there is a risk of bias pertaining to the scoring and interpretation of treatment response. There also remains an unmet need for prospective studies that compare features between models of treatment response and remission outcomes. Thus far, only one study [33] has assessed both outcomes, although it did not report a difference in top features between these models. It remains to be determined whether there are

reproducible features that are specific to reaching threshold for treatment response, relative to treatment remission.

Most studies contained in the present review (86.6%) used binary classification models to discriminate treatment responders' treatment from non-responders. Studies varied in terms of the specific clinical scale and change-score thresholds that constituted treatment response. Overall, four studies (26.6%) selected a ≥ 50% reduction on the HAMD-17 [16] as the threshold of clinical response, while three studies (20%) defined clinical response as ≥ 50% reduction on the MADRS [17]. Large differences in treatment duration were also observed across trials. Importantly, greater standardisation in how clinical response is defined is required to better assess the performance of prospective models, aid in the reproducibility of findings, and improve the likelihood of real-world clinical utility of ML models in psychiatry. Similarly, as described elsewhere [18], there is a lack of clear consensus on how treatment resistance is defined, which highlights the need for greater consistency across studies. Furthermore, only three studies (20%) assessed the performance of multiple algorithms, which limits a comparison on which algorithms tended to perform well.

Within chapter 7, in the context of classification models, it is important to highlight that uncertainty estimates should be considered when evaluating model accuracy and other common performance metrics such as sensitivity and specificity. For instance, while a specific model may show a reasonable accuracy, if a large range is observed between the upper and lower bounds of the 95% confidence interval, it is plausible that the model may be too imprecise to reasonably predict treatment response or selection in a prospective trial. Therefore, in the absence of uncertainty estimates such as confidence intervals, it is imperative that model performance is

interpreted with necessary caution. It is also worth noting the inherent difficulty in estimating the variability of cross-validated performance metrics [19]. Additionally, many other fields successfully use cross-validation as a basis for choosing between different models or tuning regularization parameters for a model, rather than taking its performance estimate at face value [20].

Within the current review, only 6 of 26 studies (23.0%) incorporated training and testing sets during model development, allowing for a comparison of uncertainty estimates across these models. Among them, only five studies (19.2%) reported either the standard deviation of model accuracy or 95% confidence intervals. As such, there remains an urgent need for prospective models to report the uncertainty estimates of performance metrics. Apart from the important considerations of uncertainty estimates, there is a need to consider the relationship between performance metrics and their implications within precision medicine. Common methods of evaluating the performance of ML classification models across studies contained within this review include accuracy, sensitivity, specificity, PPV, NPV, and AUC.

Although these metrics all provide useful information to evaluate the potential utility of the model, it is important to consider the relationship between them and their likely expected benefits for treatment selection. For instance, seventeen of twenty-six studies (65.38%) used a binary classification task to predict clinical response vs. non-response to a specific intervention. In this instance, the sensitivity of the model corresponds to its ability to correctly identify patients who will respond to the intervention (true positive), while specificity relates to the ability to identify patients who are likely to be non-responders (true negative). Additionally, PPV

and NPV provide insight into the prevalence of the outcome, and indicate the likelihood of clinical response, or non-response, in the case of a positive or negative result, respectively.

Furthermore, while the ideal threshold between sensitivity and specificity largely depends on the baseline rates of treatment efficacy for a given intervention, it is important to highlight that reasonable balanced accuracy does not necessarily translate into a model with clinical utility or scalability. For example, a binary classification model with a balanced accuracy of 67.5% in predicting response vs nonresponse to clozapine, corresponding to 45% sensitivity (true positive) and 85% specificity (true negative), shows worse performance than random chance at identifying whether a given patient will meet a pre-specified threshold for clinical response to the medication. While clozapine has been shown to be an effective treatment in psychotic disorders [21], it also facilitates a host of undesirable side effects, including drowsiness, hypersalivation, and constipation [22]. As such, this hypothetical model will perform extremely poorly in identifying which patients will respond to clozapine, and the associated predictors lack discriminative capabilities in this regard. In other words, important features, or biomarkers, within this model provide a signal for identifying whether a patient will not respond to clozapine but fail to provide meaningful signals for therapeutic response. Conversely, even with an 85% specificity (true negative), this model will misclassify patients as non-responders in 15% of cases. This misclassification error, or number of false negatives, scales proportionally to the overall sample size, leading to many individuals prescribed a medication with many adverse side effects that will ultimately be ineffective when implemented clinically.

Therefore, when evaluating performance thresholds to ascertain whether a given model is sufficiently accurate to make a useful impact in selecting treatments, it is important to consider

the expected efficacy of the intervention, the therapeutic safety profile, and whether the proportion of true positives and true negatives within a model provide a meaningful performance threshold for a given disorder and intervention. Moreover, metrics such as PPV and NPV provide useful context into the prevalence of a given outcome, and should be considered alongside sensitivity, specificity, and AUC.

9.4. Future Directions

Within chapter 2, while the HARM model showed reasonable performance, further refinement is needed in prospective models, and a much smaller error rate is required to implement such predictive models as clinical tools. Furthermore, as mentioned previously, variables with more than 15% missing data were excluded from the analysis. Other imputation strategies, such as *k*-nearest neighbours [23], and multiple imputation by chained equations (MICE) [24] may be a useful alternative in the case of missing data. Nonetheless, it is important to note that each imputation strategy has its own set of limitations [25]. Other algorithms, and pre-processing strategies, may lead to different performance metrics. Additionally, considering both the small sample size and low base rate, models developed using larger training sets, and prospective validation, are required. Within chapter 3, there is a need for models that use a wider framework when selecting input data to use as candidate features. Considering that our model performance is directly dependent on the available input data, an exploratory data-driven approach may be warranted in predictive models. Most machine learning studies in forensic psychiatry thus far focus purely on clinical and administrative data, given the widespread availability of such data. However, other modalities, such as neuroimaging (MRI, fMRI, DTI), electrophysiology (EEG, MEG, ERG) various sensors (actigraphy, heart rate variability), and genomic features (whole genome

sequencing, whole exome sequencing, and RNA sequencing) may prove to facilitate model performance, when used in conjunction with clinical data. Moreover, longitudinal studies with larger multicentric samples and adequate external validation are needed to translate proof-of-concept predictive models into applications to be used in clinical and legal settings. We hypothesize that such models may facilitate a more personalized approach to patient evaluation and risk management, provide greater precision in deriving a tailored treatment plan, and aid clinicians and the legal system in the decision-making process as it pertains to mentally disordered offenders. Ultimately, they may become critical tools to assist in prison sentencing, to determine fitness to stand trial, and to optimize the progress of individuals in the forensic system towards rehabilitation.

Within chapter 6, to facilitate EEG biomarkers of response to specific treatments, future studies may benefit from testing model performance on external datasets of other psychiatric medications or neurostimulation therapies. For example, Wu and colleagues assessed whether the algorithm SELSER, trained on SSRI datasets, could predict response to rTMS [26]. This approach may help highlight differences in important features to predict treatment response across psychiatric medications and provide an avenue to investigate potential neurophysiological mechanisms of action. Moreover, by exploring whether models retain similar features and modest prediction accuracy when tested on external datasets of other interventions, this may provide a way to identify generalizable EEG biomarkers that are related to therapeutic improvement or treatment resistance across disorders. Nonetheless, it may be more informative and realistic to focus on predictors of response to specific classes of medications and neurostimulation trials, to identify divergent mechanisms of therapeutic efficacy and treatment

resistance. Either way, this will require a careful consideration of differences in outcome instruments between datasets. As demonstrated in the current review, studies varied largely in the number of electrodes used, EEG systems, feature selection and extraction methods, and machine learning algorithms. Considering the heterogeneity observed across studies, large, standardised datasets must become available before this field can move ahead in a significant way. Importantly, there is a need for models developed using large well-characterised samples, with separate training, testing, and external validation datasets, to derive classification tools that can be useful clinically. Similarly, available repositories are needed to appropriately replicate models developed thus far, identify generalizable biomarkers of treatment response across interventions, and identify distinct neurophysiological markers that can help guide treatment selection in MDD.

In chapter 7, models of treatment response within randomized clinical trials have been developed using peripheral blood markers comprising SNPs and fatty acid composition, resting-state EEG, resting-state, and task-specific fMRI, as well as multimodal data comprising combinations of clinical, genetic, EEG, and fMRI features. Besides the approaches used in the literature thus far, there are several types of features that may be useful to incorporate in prospective models of treatment response and selection.

In terms of whole-blood peripheral biomarkers, next-generation sequencing methods such as RNA sequencing (RNA-seq) can be used to identify gene expression markers that are predictive of treatment response. For instance, Nøhr and colleagues [27] used data from a placebo-controlled trial comprising 184 patients treated with either vortioxetine or placebo for MDD, and using

blood samples collected with PAX gene tubes, identified three novel genes whose RNA expression levels at baseline and week 8 were significantly (FDR <0.05) associated with treatment response after 8 weeks of treatment. However, they did not identify any genes that were differentially expressed between placebo and vortioxetine groups [27]. More recently, new low-cost, portable high-throughput single-cell RNA sequencing methods have been developed, which have been used for cell-specific biomarker discovery [28]. Importantly, new feature selection methods are available for biomarker discovery using sparse single cell data. For example, it was shown that a probabilistic generative model can reduce the high-dimensional space in single-cell gene expression data and provide uncertainty estimates [29].

With respect to neurophysiological measures such as EEG, new multimodal techniques have been developed, such as combining TMS with EEG, to directly and non-invasively explore cortical reactivity with improved temporal resolution [30]. This allows for examining several types of features, including cortical excitability, cortical inhibition, cortical oscillations, and the balance between excitation and inhibition within the cortex in response to TMS pulses. This technique may be particularly useful in randomized trials of rTMS, by measuring baseline brain neurophysiology and mid-treatment. For instance, in a study by Voineskos and colleagues [31], N45 amplitude measured using TMS-EEG over the DLPFC was shown to discriminate individuals with depression from healthy controls with 76.6% accuracy (80% sensitivity, 73.3% specificity, AUC: 0.829) [31].

In terms of functional neuroimaging, functional near-infrared spectroscopy (fNIRS) is a method that uses near-infrared light to estimate cortical hemodynamic activity in response to neural

activity [32]. While fNIRS has several remaining limitations [33], such as a depth sensitivity of approximately 1.5 cm, and a spatial resolution up to 1 cm, it has recently been used to dichotomize patients with MDD from healthy controls, with frontal region integral values correctly classifying 75.2% of patients with MDD, and 74.3% of healthy controls, respectively [34]. However, it remains to be investigated whether this has utility in identifying predictors of treatment response between individuals within the same diagnostic category.

Furthermore, in terms of low-cost features that may be predictive of treatment response, there is increasing interest in the use of speech-based biomarkers adopted using smartphone technology [35]. For instance, in a study by Mundt and colleagues [36] comprising 105 adults with MDD, it was found that baseline and week 4 speech markers could predict responder vs non-responder status to sertraline at week 4 with a sensitivity estimate of 70.6% and specificity estimate of 79.2%, respectively. Moreover, six vocal acoustic measures were found to significantly correlate with depressive severity scores, as measured using the Quick Inventory of Depressive Symptomatology - Clinician Rating (QIDS-C) scale. This included total pause time, pause variability, percent pause time, speech/pause ratio, and speaking rate [36].

While RCTs and evidence-based medicine have facilitated undeniable advancements in patient care, personalised interventions remain a critical need in mental health [37]. Machine-learning precision trials may help us move away from the “one size fits all” assumption of current trials by including patient heterogeneity in individualized models. Similarly, assigning patients to a randomly selected dose in the established therapeutic range, while keeping important considerations such as body weight and contraindications in mind, may facilitate useful

algorithms to titrate medications with greater granularity. However, this will require large sample sizes, and appropriate training, testing, and external validation prior to clinical implementation.

Importantly, although treatment response prediction has utility in prognosticating whether a patient will respond to a specific intervention, they cannot determine the optimal treatment option for a specific patient. As such, machine-learning guided models of treatment selection, evaluating individual differences in comparative effectiveness across the same group of patients, are required to facilitate precision psychiatry.

9.5. Conclusion

The results of this thesis have advanced the field of precision psychiatry by 1) developing a predictive model of longitudinal physical aggression in patients with schizophrenia in forensic settings and comparing such models against clinician judgement alone, and models that combine data-driven approaches and clinician insight. Here, we showed that clinical judgement alone shows a high false negative rate, where patients who will commit physical aggression are incorrectly identified as low risk. Conversely, the data-driven HARM models showed a low false negative rate, correctly identifying the majority of patients who are low risk. However, clinical judgement showed a higher positive predictive value than the data-driven HARM models, suggesting that such prognostic tools may have utility as an adjunct to clinical judgement; 2) systematically synthesizing existing studies on predicting criminal and violent outcomes in psychiatry, and confirming the utility of evidence-based risk factors in developing models with an average sensitivity of 73.33% (95% CI: 64.09-79.63) and average specificity of 72.90% (95% CI: 63.98-79.66), respectively , 3) highlighting the importance of considering ethical constraints in models within forensic psychiatry, and 4) systematically synthesizing existing studies on

predicting treatment response using EEG in MDD, and in the context of clinical trials more broadly, as well as identify methodological recommendations for prospective machine-learning guided trials.

9.6. References

1. Grann, M., Belfrage, H. & Tengström, A. Actuarial assessment of risk for violence: Predictive validity of the VRAG and the historical part of the HCR-20. *Crim Justice Behav* (2000) doi:10.1177/0093854800027001006.
2. Dernevik, M., Grann, M. & Johansson, S. Violent behaviour in forensic psychiatric patients: Risk assessment and different risk-management levels using the HCR-20. *Psychology, Crime and Law* (2002) doi:10.1080/10683160208401811.
3. Sharma, P. K. *et al.* Issues and challenges of data security in a cloud computing environment. *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017* **2018-Janua**, 560–566 (2017).
4. Warnat-Herresthal, S. *et al.* Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
5. Jansman-Hart, E. M., Seto, M. C., Crocker, A. G., Nicholls, T. L. & Côté, G. International Trends in Demand for Forensic Mental Health Services. *Int J Forensic Ment Health* **10**, 326–336 (2011).
6. Phillips, J. L. *et al.* Single , Repeated , and Maintenance Ketamine Infusions for Treatment-Resistant Depression : A Randomized Controlled Trial. *Am J Psych* **176**, 401–409 (2019).
7. Medeiros, G. C. *et al.* Blood-based biomarkers of antidepressant response to ketamine and esketamine: A systematic review and meta-analysis. *Mol Psychiatry* **27**, 3658–3669 (2022).
8. Matveychuk, D. *et al.* Ketamine as an antidepressant: overview of its mechanisms of action and potential predictive biomarkers. *Ther Adv Psychopharmacol* **10**, 204512532091665 (2020).
9. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip Rev Comput Stat* **2**, 433–459 (2010).

10. Friedman, J. H. & Roosen, C. B. An introduction to multivariate adaptive regression splines. *Stat Methods Med Res* **4**, 197–217 (1995).

11. Fitkov-Norris, E., Vahid, S. & Hand, C. Evaluating the Impact of Categorical Data Encoding and Scaling on Neural Network Classification Performance: The Case of Repeat Consumption of Identical Cultural Goods. in *Communications in Computer and Information Science* vol. 311 343–0352 (2012).

12. Dash, M. & Liu, H. Feature selection for classification. *Intelligent Data Analysis* (1997) doi:10.3233/IDA-1997-1302.

13. Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. in *Data Classification: Algorithms and Applications* (2014). doi:10.1201/b17320.

14. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Computers and Electrical Engineering* (2014) doi:10.1016/j.compeleceng.2013.11.024.

15. Rosellini, A. J. *et al.* Predicting non-familial major physical violent crime perpetration in the US Army from administrative data. *Psychol Med* (2016) doi:10.1017/S0033291715001774.

16. Zimmerman, M., Martinez, J. H., Young, D., Chelminski, I. & Dalrymple, K. Severity classification on the Hamilton depression rating scale. *J Affect Disord* **150**, 384–388 (2013).

17. Quilty, L. C. *et al.* The structure of the Montgomery-Åsberg depression rating scale over the course of treatment for depression. *Int J Methods Psychiatr Res* **22**, 175–184 (2013).

18. Howes, O. D., Thase, M. E. & Pillinger, T. Treatment resistance in psychiatry: state of the art and new directions. *Molecular Psychiatry* Preprint at https://doi.org/10.1038/s41380-021-01200-3 (2021).

19. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* vol. 180 68–77 Preprint at https://doi.org/10.1016/j.neuroimage.2017.06.061 (2018).

20. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning with Applications in R Second Edition*. (2021).

21. Wagner, E. *et al.* Efficacy and safety of clozapine in psychotic disorders—a systematic quantitative meta-review. *Translational Psychiatry* vol. 11 Preprint at https://doi.org/10.1038/s41398-021-01613-2 (2021).

22. Farooq, S. & Taylor, M. Clozapine: Dangerous orphan or neglected friend? *British Journal of Psychiatry* vol. 198 247–249 Preprint at https://doi.org/10.1192/bjp.bp.110.088690 (2011).

23. Zhang, Z. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med* **4**, 218–218 (2016).

24. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* **20**, 40–49 (2011).

25. Poulos, J. & Valle, R. *MISSING DATA IMPUTATION FOR SUPERVISED LEARNING †*. (2018).

26. Wu, W. *et al.* An electroencephalographic signature predicts antidepressant response in major depression. *Nat Biotechnol* **38**, 439–447 (2020).

27. Nøhr, A. K. *et al.* A large-scale genome-wide gene expression analysis in peripheral blood identifies very few differentially expressed genes related to antidepressant treatment and response in patients with major depressive disorder. *Neuropsychopharmacology* **46**, 1324–1332 (2021).

28. Gierahn, T. M. *et al.* Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nat Methods* **14**, 395–398 (2017).

29. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat Commun* **9**, (2018).

30. Hill, A. T., Rogasch, N. C., Fitzgerald, P. B. & Hoy, K. E. TMS-EEG: A window into the neurophysiological effects of transcranial electrical stimulation in non-motor brain regions. *Neuroscience and Biobehavioral Reviews* Preprint at https://doi.org/10.1016/j.neubiorev.2016.03.006 (2016).

31. Voineskos, D. *et al.* Altered Transcranial Magnetic Stimulation–Electroencephalographic Markers of Inhibition and Excitation in the Dorsolateral Prefrontal Cortex in Major Depressive Disorder. *Biol Psychiatry* **85**, 477–486 (2019).

32. Ferrari, M. & Quaresima, V. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage* vol. 63 921–935 Preprint at https://doi.org/10.1016/j.neuroimage.2012.03.049 (2012).

33. Chen, W. L. *et al.* Functional Near-Infrared Spectroscopy and Its Clinical Application in the Field of Neuroscience: Advances and Future Directions. *Frontiers in Neuroscience* vol. 14 Preprint at https://doi.org/10.3389/fnins.2020.00724 (2020).

34. Husain, S. F. *et al.* Validating a functional near-infrared spectroscopy diagnostic paradigm for Major Depressive Disorder. *Sci Rep* **10**, (2020).

35. Fagherazzi, G., Fischer, A., Ismael, M. & Despotovic, V. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers* vol. 5 78–88 Preprint at https://doi.org/10.1159/000515346 (2021).

36. Mundt, J. C., Vogel, A. P., Feltner, D. E. & Lenderking, W. R. Vocal acoustic biomarkers of depression severity and treatment response. *Biol Psychiatry* **72**, 580–587 (2012).

37. Cavalcante, I. & Benson, P. Machine learning-guided intervention trials to predict treatment response at an individual patient level : an important second step following randomized clinical trials. *Mol Psychiatry* 701–702 (2020) doi:10.1038/s41380-018-0250-y.