UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM PSIQUIATRIA E CIÊNCIAS DO COMPORTAMENTO

Diego Librenza Garcia

PSIQUIATRIA PREDITIVA E PERSONALIZADA: APLICAÇÕES DE TÉCNICAS DE MACHINE LEARNING EM SAÚDE MENTAL

Porto Alegre, 2019

DIEGO LIBRENZA GARCIA

PSIQUIATRIA PREDITIVA E PERSONALIZADA: APLICAÇÕES DE TÉCNICAS DE MACHINE LEARNING EM SAÚDE MENTAL

Tese apresentada como requisito parcial para obtenção de título de Doutor em Psiquiatria à Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Psiquiatria e Ciências do Comportamento.

Orientador: Prof. Ives Cavalcante Passos

Porto Alegre, 2019

FOLHA DE APROVAÇÃO DA BANCA EXAMINADORA

DIEGO LIBRENZA GARCIA

PSIQUIATRIA PREDITIVA E PERSONALIZADA: APLICAÇÕES DE TÉCNICAS DE MACHINE LEARNING EM SAÚDE MENTAL

Dissertação apresentada como requisito parcial para obtenção de título de Doutor em Psiquiatria à Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Psiquiatria e Ciências do Comportamento.

Porto Alegre, 13 de setembro de 2019.

A comissão Examinadora, abaixo assinada, aprova a tese "Psiquiatria preditiva e personalizada: aplicações de técnicas de *machine learning* em saúde mental", elaborada por Diego Librenza Garcia, como requisito parcial para a obtenção do grau de Doutor em Psiquiatria e Ciências do Comportamento.

_____

Prof. Dr. Felix Henrique Paim Kessler (UFRGS)

_____

Prof. Dr. Analuiza Camozzato de Pádua (UFCSPA)

_____

Prof. Dr. Ygor Arzeno Ferrão (UFCSPA)

_____

Prof. Dr. Ives Cavalcante Passos - Orientador

“POZZO:
I am blind.
(Silence.)
ESTRAGON:
Perhaps he can see into the future.”
— *Samuel Beckett, Waiting for Godot*

“This is only a foretaste of what is to come,
and only the shadow of what is going to be.”
— *Alan Turing, Times, 1949*

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Ives Cavalcante Passos, pelos inúmeros ensinamentos e oportunidades que recebi, cruciais para minha carreira e meu desenvolvimento como pesquisador.

Ao professor Flavio Kapczinski, por ter sido determinante na minha escolha pela psiquiatria e neurociência, e por todo o apoio e conhecimento que recebi ao longo dos anos.

A todos os professores que fizeram parte de minha formação, aumentando minha paixão pela psiquiatria clínica e pela pesquisa, e me tornando um melhor profissional. Em especial, agradeço aos professores Ana Luiza Camozzato de Pádua, Felix Henrique Paim Kessler, Neusa Sica da Rocha, e Ygor Arzeno Ferrão, e aos doutores Ariel Roitman, Gledis Lisiane Correa Luz Motta, e Madeleine Scop Medeiros.

Ao professor André Russowsky Brunoni e à equipe do ELSA-Brasil pela oportunidade de colaborar neste trabalho.

À Universidade Federal do Rio Grande do Sul, presente e determinante em minha vida em três momentos distintos: o bacharelado em física, a graduação em medicina, e neste doutorado.

À McMaster University, por ter me acolhido em meu fellowship, desde 2018.

À minha família, pelo constante apoio e carinho que recebi durante essa trajetória. Em especial, ao meu pai, Roberto Leite Garcia, a quem eu devo tudo que alcancei em minha carreira, por sempre priorizar a mim e à minha educação.

RESUMO

Técnicas de *machine learning* ganharam tração nos últimos anos devido não só à exponencial evolução na capacidade de processamento dos computadores, mas também pela sua habilidade de solucionar problemas de alta complexidade. Recentemente, estas técnicas começaram a ser usadas para abordar limitações no campo da psiquiatria e da saúde mental, na tentativa de desenvolver abordagens mais personalizadas e que integrem diferentes níveis de conhecimento. Em nosso primeiro artigo, revisamos aplicações de *machine learning* no transtorno de humor bipolar, incluindo classificação diagnóstica, predição de desfechos desfavoráveis, como suicídio, recorrência de episódios, e resposta ao tratamento, e a busca por fenótipos *data-driven* para o transtorno. Foram incluídos um total de 51 artigos nesta revisão sistemática. No mesmo estudo, conduzimos uma meta-análise de estudos que usaram neuroimagem para diferenciar transtorno de humor bipolar de controles saudáveis, obtendo áreas sob a curva (AUC) de 0.698, 0.754, e 0.712, para ressonância magnética estrutural, funcional, e ambas combinadas, respectivamente. No segundo artigo, focamos nas aplicações de *machine learning* para predizer resposta ao tratamento farmacológico e não-farmacológico em transtornos psiquiátricos. Foram incluídos 61 estudos, correspondendo a um *pool* de 46,957 pacientes, e nós discutimos o uso de diferentes níveis de dados para predizer resposta a tratamento e as limitações metodológicas desses estudos.

Um dos grandes desafios em psiquiatria é a heterogeneidade que os transtornos psiquiátricos tem em sua apresentação e trajetória. Embora resultados a níveis de grupo sejam conhecidos, há uma escassez de dados na literatura sobre trajetórias individuais destes transtornos, fato que cria obstáculos para intervenções mais precoces e personalizadas. Para tentar suprir essa lacuna, nós conduzimos uma análise de *machine learning* usando dados do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil), uma coorte ocupacional com 15,105 pacientes avaliados entre 2008-2010 (onda 1) e reavaliados entre 2012-2014 (onda 2). Foram incluídos no ELSA-Brasil funcionários de seis instituições públicas de ensino brasileiras entre 35 e 74 anos de idade. Nós definimos três trajetórias distintas: (1) indivíduos sem depressão em ambas ondas 1 e 2;

(2) indivíduos com depressão apenas na onda 2 (depressão incidente); e (3) pacientes com depressão em ambas as ondas (depressão persistente). Usando preditores clínicos e sociodemográficos, nós desenvolvemos três modelos de *machine learning* com *embedded feature selection* usando o algoritmo *Elastic Net*: (1) para diferenciar indivíduos com depressão daqueles sem depressão; (2) para diferenciar indivíduos com depressão incidente daqueles sem depressão; e (3) para diferenciar pacientes com depressão persistente daqueles sem depressão. Foram obtidos para os três modelos AUCs de 0.90 (95% CI 0.85 - 0.95), 0.89 (95% CI 0.85 - 0.94), e 0.90 (95% CI 0.86 - 0.95), respectivamente. Variáveis clínicas, como presença de comorbidades psiquiátricas, foram consistentemente mais relevantes para os modelos preditivos, sobretudo o transtorno de ansiedade generalizada e o transtorno obsessivo-compulsivo, que estiveram entre as três variáveis mais relevantes em todos os modelos. Dentre as variáveis sociodemográficas, a mais relevante foi o sexo biológico. Estes resultados demonstram que é possível prever o diagnóstico e o prognóstico de depressão em nível individual integrando variáveis clínicas e sociodemográficas.

Palavras-chave: *machine learning*, modelos preditivos, *big data*, transtorno bipolar, transtorno depressivo, meta-análise, diagnóstico, prognóstico, tratamento personalizado.

ABSTRACT

Machine learning techniques have gained traction in the last few years due not only to the exponential advances in the processing power of computers but also for its ability to solve highly complex problems. Recently, these techniques have been used to address longstanding limitations in the field of psychiatry and mental health, in an attempt to develop personalized approaches that can integrate different fields of knowledge. In our first publication, we reviewed applications of machine learning in the study of bipolar disorder, including diagnostic classification, prediction of poor outcomes, such as suicide, mood episode recurrence, and treatment response, as well as the search for data-driven phenotypes of bipolar disorder. We included 51 papers in this systematic review. In the same study, we performed a meta-analysis of studies that used neuroimaging to differentiate bipolar disorder from healthy controls, obtaining areas under the curve (AUC) of 0.698, 0.754, and 0.712 for functional magnetic resonance, structural magnetic resonance, and both combined, respectively. In the second paper, we focused on the applications of machine learning to predict response to the pharmacologic and non-pharmacologic treatment of psychiatric disorders. We included 61 studies, with a pool of 46,957 patients, and discussed the use of different levels of data to predict treatment response, and the methodological limitations of these studies.

One critical challenge in psychiatry is the heterogeneity of the presentation and trajectory of psychiatric disorders. Although results at the group-level are known, there is a scarcity of data in the literature about the individual trajectory of these disorders, which creates obstacles for more timely and personalized interventions. To address this gap, we performed a machine learning analysis in the Longitudinal Study of Adult Health (ELSA-Brasil), an occupational cohort with 15,105 patients assessed between 2008-2010 (wave 1), and reassessed between 2012-2014 (wave 2). Public servants between the ages of 35 and 74 years of six Brazilian institutes were included in the ELSA-Brasil cohort. We defined three distinct trajectories: (1) subjects without depression at both wave 1 and 2; (2) subjects with depression only at wave 2 (incident depression); and (3) subjects with depression at both wave 1 and 2 (persistent depression). Using clinical and sociodemographic predictors, we developed three machine learning models with

embedded feature selection using the Elastic Net algorithm: (1) to differentiate subjects with depression from those without depression; (2) to differentiate subjects with incident depression from those without depression; and (3) to differentiate subjects with persistent depression from those without depression. We obtained AUCs of 0.90 (95% CI 0.85 - 0.95), 0.89 (95% CI 0.85 - 0.94), and 0.90 (95% CI 0.86 - 0.95, respectively). Clinical variables, such as comorbidities, were consistently more relevant for the predictive models. Generalized anxiety disorder and obsessive-compulsive disorder, especially, were among the three most relevant features in all models. Among the sociodemographic features, sex was the most relevant variable. Our results demonstrate that it is possible to predict the diagnosis and prognosis of depression at an individual-level by integrating clinical and sociodemographic variables.

Keywords: machine learning, predictive models, big data, bipolar disorder, depressive disorder, meta-analysis, diagnosis, prognosis, personalized treatment.

LISTA DE ABREVIATURAS E SIGLAS

AAP *Atypical Antipsychotic* – Antipsicótico atípico

ACC *Anterior cingulate cortex* – Córtex cingulado anterior

AD *Alzheimer's Disease* – Doença de Alzheimer

ADHD *Attention deficit hyperactivity disorder* – Transtorno do déficit de atenção/hiperatividade

ANN *Artificial neural networks* – Rede neural artificial

AOD *Auditory oddball*

ARMS *At-risk mental states* – Estados mentais de risco

ASL *Arterial spin labelling*

AUC *Area under the curve* – Área sob a curva

AUD *Alcohol use disorder* – Transtorno por uso de álcool

BD *Bipolar disorder* – Transtorno bipolar

BN *Bayesian networks* – Redes bayesianas

BOLD *Blood-oxygen-level dependent imaging*

CANTAB *Cambridge Neurocognitive Test Automated Battery* – Bateria Computadorizada de Testes Neuropsicológicos de Cambridge

CART *Classification and regression trees* – Árvore de classificação e regressão

CBT *Cognitive behavioral therapy* – Terapia cognitiva comportamental

CCR *Correct classification rate* – Taxa de classificação correta

CID-11 Classificação estatística internacional de doenças e problemas relacionados à saúde, 11a versão

CIS-R *Clinical Interview Schedule-Revised*

DBM *Deformation-based morphometry* – Morfometria baseada em deformação

DSM-V *Diagnostic and statistical manual of mental disorders, 5th edition* – Manual diagnóstico e estatístico de transtornos mentais, 5ª edição

DTI *Diffusion Tensor Imaging* – Imagem de tensor de difusão

DWI *Diffusion-weighted Imaging* – Imagem ponderada de difusão

EBM *Evidence-based medicine* – Medicina baseada em evidências

ECG *Electrocardiogram* – Eletrocardiograma

ECT *Eletroconvulsive therapy* – Electroconvulsoterapia

EEG *Electroencephalography* – Eletroencefalograma

ELSA Estudo Longitudinal de Saúde do Adulto

EMD *Empirical mode decomposition* – Decomposição de modo empírico

FA *Fractional anisotropy – Anisotropia fracionada*

fALFF *Fractional amplitude of low-frequency fluctuation* – Amplitude fracionada de flutuação de baixa frequência

fMRI *Functional magnetic resonance imaging* – Ressonância magnética funcional

GAD *Generalized Anxiety Disorder* – Transtorno de ansiedade generalizada

GAF *Global Assessment of Functioning* – Avaliação global de funcionamento

GM *Gray matter* – Substância cinzenta

GPC *Gaussian Process Classifier* – Classificador por processo gaussiano

HAMD-17 *17-item Hamilton Rating Scale for Depression* – Escala de Avaliação de Depressão de Hamilton, versão com 17 itens

HC *Healthy controls* – Controles saudáveis

IACO *Improved Ant Colony Optimization* – Otimização da colônia de formigas melhorado

IFITM3 *Interferon Induced Transmembrane Protein 3* – Proteína transmembrana induzida por interferon 3

LASSO *Least absolute shrinkage and selection operator* – Operador de Menor encolhimento absoluto e seleção

LOBD *Late-onset bipolar disorder* – Transtorno bipolar de início tardio

LR *Logistic regression* – Regressão logística

MAO *Monoamine oxidase* – Monoamina oxidase

MBE Medicina baseada em evidências

MDD *Major depressive disorder* – Transtorno depressivo maior

MDR *Multifactorial dimensionality reduction* – Redução dimensional multifatorial

MFA *Mixture of factor analysis*

MPH *Methylphenidate* – Metilfenidato

MRI *Magnetic resonance imaging* – Ressonância magnética

NB *Naïve Bayes*

NOS *Not otherwise specified* – Sem outra especificação

NPV *Negative predictive value* – Valor preditivo negativo

OCD *Obssessive compulsive disorder* – Transtorno obsessivo-compulsivo

PCA *Principal component analysis* – Análise de Componentes Principais

PD *Panic Disorder* – Transtorno de pânico

PSO *Particle swarm optimization* – Otimização por enxame de partículas

PPV *Positive predictive value* – Valor preditivo positivo

QCA *Quantitative Clinical Assessment* – Avaliação clínica quantitativa

QEEG *Quantitative Electroencephalogram* – Eletroencefalograma quantitativo

QIDS *Quick Inventory of Depressive Symptomatology* – Inventário rápido de sintomatologia depressiva

RCT *Randomized Clinical Trial* – Ensaio Clínico Randomizado

RdOC *Research Domain Criteria*

RF *Random Forest* – Floresta aleatória

RFE *Recursive Feature Elimination* – Eliminação recursiva de variáveis

RR *Interbeat interval series* – Intervalo entre batimentos sucessivos

rs-fMRI *Resting-state Functional Magnetic Resonance Imaging* – Ressonância magnética funcional em estado de repouso

rTMS *Repetitive Transcranial Magnetic Stimulation* – Estimulação magnética transcraniana repetitiva

RVM *Relevance vector machine* – Máquina de vetor de relevância

SAD *Social Anxiety Disorder* – Transtorno de ansiedade social

SADD *Schizoaffective disorder with depressive episodes* – Transtorno esquizoafetivo com episódios depressivos

SADM *Schizoaffective disorder with manic episodes* – Transtorno esquizoafetivo com episódios maníacos

SLC4A4 *Solute Carrier Family 4 Member 4 – Carreador Solúvel Família 4 Membro 4*

sMRI *Structural magnetic resonance imaging* – Ressonância magnética estrutural

SNP *Single-nucleotide polymorphisms* – Polimorfismo de nucleotídeo único

SROC *Summary Receiver Operating Characteristic* – Característica de Operação do Receptor Sumária

SSRI *Selective serotonin reuptake inhibitor* – Inibidores seletivos da recaptação da serotonina

STFT *Short-time Fourier Transform* – Transformada de Fourier de Curto Termo

SUD *Substance use disorder* – Transtorno por uso de substância

SVM *Support Vector Machine* – Máquina de vetores de suporte

SVM-FoBa *Support Vector Machine with forward-backward search strategy* – Máquina de vetores de suporte com estratégia de busca *forward-backward*

TCC Terapia cognitivo-comportamental

tDCS *Transcortical Direct Current Stimulation* – Estimulação transcraniana por corrente contínua

TIMP1 *Tissue Inhibitor of Metalloproteinases* – Inibidor tecidual de metaloproteinases

VBM *Voxel-based morphometry* – Morfometria baseada em voxel

WM *White matter* – Substância branca

WT *Wavelet-based technique* – Técnica baseada em *wavelet*

SUMÁRIO

1. APRESENTAÇÃO

Esta tese, intitulada "Psiquiatria Preditiva e Personalizada: Aplicações de técnicas de *machine learning* em saúde mental", foi apresentada ao Programa de Pós-Graduação em Psiquiatria e Ciências do Comportamento da Universidade Federal do Rio Grande do Sul em 13 de setembro de 2019.

Apesar dos avanços proporcionados pela medicina baseada em evidências, sobretudo na determinação de fatores de risco e efeitos de tratamento a nível de grupo, críticas limitações permanecem na prática psiquiátrica e em saúde mental. O tratamento farmacológico, por exemplo, apesar de todas as evidências disponíveis, permanece um exercício de tentativa e erro. Ademais, não é possível prever quais sujeitos vão desenvolver um transtorno psiquiátrico, ou após desenvolver, quais terão um pior prognóstico, incluindo tentativas de suicídio ou refratariedade ao tratamento. Em suma, nosso conhecimento em psiquiatria se concentra na manifestação coletiva de sintomas e desfechos, com escassas ferramentas que possam ser aplicadas com acurácia para auxiliar no curso clínico de cada indivíduo.

Machine learning é um campo do estudo de inteligência artificial focado em algoritmos computacionais que possuem a habilidade de aprender com dados e melhorar sua performance, sem serem explicitamente programados para tanto, e possui um amplo campo de aplicação, que vai desde a criação de modelos preditivos até à busca de padrões e representações dentro de dados não-estruturados. Divide-se, de forma geral, em dois ramos principais: aprendizado supervisionado e não-supervisionado. No aprendizado supervisionado, o algoritmo usa um banco de dados com desfecho conhecido e treina para detectar este desfecho. Após, é fornecido um novo banco de dados ao algoritmo sem informar o desfecho, e predições são feitas com base nos padrões aprendidos durante a fase de treino. No aprendizado não-supervisionado, não há um desfecho definido, e o algoritmo busca padrões que possam ser usados para agrupar estes dados em diferentes *clusters* de informação.

A presente tese se baseou na hipótese de que técnicas de *machine learning* têm o potencial de transformar o estudo dos transtornos psiquiátricos e melhorar desfechos clínicos. No primeiro artigo, revisamos todas as aplicações da técnica em pacientes com transtorno de humor bipolar. Este artigo foi submetido, aceito e publicado na *Neuroscience & Biobehavioral Reviews* (fator de impacto: 8.299) sob o título **The impact of machine learning techniques in the study of bipolar disorder: a systematic**

review. No segundo artigo, revisamos o uso de *machine learning* para predição do tratamento em transtornos psiquiátricos. Este artigo foi revisado na *Molecular Psychiatry* (fator de impacto: 11.973) e será submetido novamente à mesma revista após ser reescrito de acordo com as sugestões dos revisores, sob o título de **Machine learning guided intervention trials in mental health: a systematic review and methodological recommendations**. Finalmente, no terceiro artigo, desenvolvemos um modelo de *machine learning* para predizer o curso do transtorno depressivo em uma grande coorte ocupacional. Este artigo está submetido e atualmente em revisão na *Neuropsychopharmacology* (fator de impacto: 7.160) sob o título de **Prediction of depression incidence, remission, and persistence in a large occupational cohort using machine learning techniques: an analysis of the ELSA-Brasil study.**

2. INTRODUÇÃO

2.1. *Machine learning*, *big data* e medicina baseada em evidência

A medicina baseada em evidência (MBE), através do uso de métodos estatísticos tradicionais, nos ajudou a entender melhor os fatores de risco, tratamentos e prognósticos em psiquiatria (1). Os estudos padrão-ouro da MBE são o ensaio clínico randomizado e a metanálise, que, primariamente, nos fornecem resultados a nível de grupo (2). Recentemente, *Greenhalg* e colegas chamaram a atenção para algumas limitações destes estudos. Em primeiro lugar, resultados estatisticamente significativos não necessariamente representam um benefício clínico real. Segundo, os ensaios clínicos não caracterizam o perfil de multimorbidade dos pacientes em cenários reais, bem como da heterogeneidade inerente ao diagnóstico, tratamento e prognóstico em psiquiatria (3). Visto que os transtornos psiquiátricos, e em especial os transtornos de humor, possuem significativa heterogeneidade clínica, existe uma necessidade de ferramentas que possam melhor caracterizar estas populações e produzir resultados não só a nível de grupo, mas também individual (4–6).

Big data é o termo usado para caracterizar bancos de dados que contêm um extenso volume de dados, criados a uma elevada velocidade, e contendo uma variedade de formatos, que vão desde variáveis moleculares, como genômica, proteômica e metabolômica, até dados clínicos, sociodemográficos, administrativos, ambientais, e mesmo informação proveniente de mídias sociais (7–9). Dada a complexidade destes massivos bancos de dados, métodos estatísticos tradicionais não são capazes de lidar com essa elevada complexidade. *Data science* (Ciência de dados) é um campo multidisciplinar focado no uso de processos, algoritmos, métodos e sistemas que visam extrair valor e gerar ferramentas aplicáveis a partir de dados estruturados e não estruturados (10). Um dos principais métodos em *data science* para análise de *big data* é o uso de técnicas de *machine learning* (11).

Machine learning é o estudo de algoritmos computacionais que possuem a capacidade de aprender com os dados que lhe são fornecidos, extraindo padrões relevantes que possam ou ser usados para inferência em novos casos, ou para descobrir *clusters* relevantes dentro desta amostra (11,12). Assim, o estudo de *machine learning* se divide em dois ramos principais: aprendizado supervisionado, onde o desfecho é conhecido e o algoritmo é treinado para detectar este desfecho, e aprendizado não-supervisionado, método usado para encontrar padrões dentro de dados não estruturados

que possam ser usados para estabelecer subgrupos dentro destes dados. Em outras palavras, no aprendizado supervisionado nós temos um dado conjunto de dados com desfecho conhecido (*training dataset*), e o algoritmo desenvolve um modelo matemático que é capaz de predizer este desfecho para novos casos (*testing dataset*). Comparando o desfecho predito pelo algoritmo com o desfecho conhecido da amostra (previamente, ou após seguimento dos pacientes), medidas de performance deste algoritmo, como acurácia ou AUC, podem ser obtidas. Com esta técnica, é possível predizer diagnóstico e prognóstico, oportunizando a prevenção de desfechos e a instituição de intervenções individualizadas. No caso do aprendizado não-supervisionado, não há um desfecho definido, e o algoritmo irá procurar padrões que permitam dividir a amostra em subgrupos que diferem entre si, mas que possuem homogeneidade interna. O uso desta técnica é especialmente útil na determinação de fenótipos relevantes em psiquiatria (13). Em função das limitações citadas anteriormente a respeito dos métodos atuais usados em psiquiatria, técnicas de *machine learning* tem ganho expressão na psiquiatria e na medicina como um todo, dado o potencial de auxiliar no desenvolvimento de intervenções psiquiátricas personalizadas, centradas no indivíduo (12,14).

2.2. Potenciais aplicações de *machine learning* em psiquiatria e saúde mental

Existem inúmeros desafios em psiquiatria e saúde mental que carecem de solução, tanto para melhor alocação de recursos pelos sistemas de saúde, quanto para promover intervenções mais precoces que possam evitar prejuízos associados aos transtornos psiquiátricos a curto e longo prazo. Técnicas de *machine learning* podem ser usadas, entre outros: 1) como uma ferramenta diagnóstica, tanto usando os diagnósticos já estabelecidos pelo DSM-V e CID-11, quanto para a descoberta de novos diagnósticos *data-driven* (13); 2) para prevenção de resposta ao tratamento, tanto farmacológico quanto não farmacológico, e de efeitos colaterais (15); e 3) para predição de desfechos desfavoráveis, como suicídio, recorrência de episódios, neuroprogressão, ou prejuízos funcionais (16,17).

Mesmo com mais de um século de estudo em psiquiatria, o diagnóstico permanece sendo feito com base em sinais e sintomas, ignorando o complexo processo patofisiológico que culmina no transtorno e em suas diferentes apresentações (18). Estes incluem dados moleculares, genéticos, de neuroimagem, comportamentais, dentre outros. Ademais, mesmo em ambientes de pesquisa, há dificuldade em integrar estes

diferentes níveis de dados. *Machine learning* pode ser usado para classificar e diferenciar transtornos psiquiátricos com base nestas informações (7). Por exemplo, a distinção entre depressão unipolar e depressão bipolar é crucial para o prognóstico, visto que estas condições possuem tratamentos distintos, e a incorreta classificação pode levar à instituição de um tratamento com potencial de piorar o curso e a apresentação da doença, ou de adicionar efeitos colaterais de uma medicação ineficaz (16,19). O uso de técnicas de neuroimagem, eletroencefalograma, e marcadores séricos, por exemplo, podem ser explorados na busca de uma diferenciação mais objetiva entre classes diagnósticas, fundamentada nas alterações neurofisiológicas encontradas nos indivíduos. Por outro lado, o aprendizado não-supervisionado também pode ser usado para descobrir *clusters* diagnósticos mais relevantes. Fazendo uso de múltiplos níveis biológicos, um algoritmo poderia encontrar novos grupos de crescente complexidade, mas também aplicabilidade, distintos daqueles sugeridos pelos manuais diagnósticos (13).

Outra necessidade em psiquiatria e saúde mental é saber qual tratamento é o ideal para cada indivíduo. *Machine learning* pode ser usado para prever quais pacientes irão responder a uma medicação, ou que pacientes vão desenvolver um efeito colateral severo, ajudando a indicar e evitar a prescrição de certos medicamentos (15). Como ensaios clínicos em geral excluem pacientes com multimorbidades para maximizar o tamanho de efeito e diminuir o número de confundidores, o tratamento na clínica acaba sendo muito diferente daquele visto em pesquisa. Desta forma, *guidelines* e metanálises não são capazes de mapear a complexidade dos pacientes "reais" (3). Além disso, saber quais indivíduos responderão a tratamentos como terapia cognitiva comportamental (20,21) e eletroconvulsoterapia (22) pode otimizar o uso de recursos em saúde pública. Ao evitar a indicação de um paciente que não vai responder à TCC, abrem-se mais vagas para aqueles que irão se beneficiar do tratamento; ao evitar a indicação de um paciente que não será responsivo à ECT, estamos evitando não só gastos desnecessários, mas também a indução de efeitos colaterais importantes provenientes de um tratamento que não seria efetivo.

Por último, *machine learning* ainda pode ser usado para predição do curso dos transtornos psiquiátricos e de seus desfechos desfavoráveis. As possibilidades são muitas: predizer transtornos psiquiátricos antes deles se manifestarem; predizer novos episódios de humor; predizer tentativas de suicídio antes que elas ocorram; predizer quais sujeitos terão depressão persistente e quais entrarão em remissão; e assim por

diante. Prever estes eventos seria um grande passo em direção à uma psiquiatria preventiva (12,23,24). Os impactos destas técnicas podem ser profundos para a sociedade, diminuindo gastos com auxílio benefício e tratamentos a nível terciário, como internações, e também para os indivíduos, evitando prejuízos funcionais e cognitivos, e promovendo uma melhor qualidade vida.

3. JUSTIFICATIVA

Embora a medicina baseada em evidência e métodos estatísticos tradicionais tenham contribuído para o avanço da psiquiatria, existem limitações para ambas as técnicas. Efeitos estatisticamente significativos não necessariamente se traduzem em efeitos clínicos relevantes, enquanto ensaios clínicos falham em contemplar a heterogeneidade dos transtornos psiquiátricos, especialmente o perfil de multimorbidade associado a estes quadros. Ademais, o volume de informação disponível, incluindo estudos e *guidelines*, se tornou massivo demais para ser manejado de maneira prática por profissionais de saúde, sendo que descobertas em pesquisa raramente são traduzidos para a prática clínica. Existe, portanto, a necessidade de avançar para além dos resultados de grupo, em direção a uma abordagem mais individualizada e personalizada. Por fim, métodos estatísticos tradicionais não são capazes analisar a complexidade de extensos bancos de dados, em especial os não-estruturados. Técnicas de *machine learning*, por outro lado, são ideais para abordar estes desafios.

Diante disso, conduzimos duas revisões sistemáticas e um estudo original com foco no uso de *machine learning* em saúde mental. Na primeira revisão e metanálise, estudamos todas as possíveis aplicações do método no estudo do transtorno de humor bipolar, e a acurácia do uso de neuroimagem para este diagnóstico. Na segunda revisão, focamos em uma dessas aplicações: a predição de resposta a tratamento em transtornos psiquiátricos. No terceiro estudo, nós aplicamos estas técnicas em uma grande coorte para predizer o curso do transtorno depressivo, incluindo predição de depressão, depressão incidente, e depressão persistente.

4. OBJETIVOS

4.1. Objetivo geral

Determinar aplicações de técnicas de *machine learning* em psiquiatria e saúde mental.

4.2. Objetivos específicos

a) Revisar e metanalisar a literatura relacionada com aplicações de técnicas de *machine learning* no transtorno de humor bipolar.

b) Revisar a literatura relacionada com o uso de técnicas de *machine learning* na predição de resposta ao tratamento em estudos com intervenções para o tratamento de transtornos psiquiátricos.

c) Utilizar técnicas de *machine learning* para predizer o curso do transtorno depressivo em uma extensa coorte ocupacional.

5. CONSIDERAÇÕES ÉTICAS

O Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil) é uma coorte multicêntrica composta por mais de 15 mil funcionários públicos de seis instituições de ensino superior brasileiras, cujo propósito é investigar incidência e fatores de risco para doenças crônicas.

O protocolo do estudo atendeu à Resolução 196/96a e a outras complementares – a Resolução CNS 346/05 Projetos multicêntricos e a Resolução CNS 347/05 Armazenamento de materiais biológicos. Foi aprovado nos comitês de ética em pesquisa das instituições envolvidas e na Comissão Nacional de Ética em Pesquisa do Conselho Nacional de Saúde (Conep). Conta ainda com o Comitê de Ética, Recrutamento e Comunicação Social, responsável por assessorar e coordenar o cumprimento de aspectos éticos e de comunicação com as instituições e indivíduos participantes do estudo. Este comitê foi responsável pela consolidação do protocolo e do Termo de Consentimento Livre e Esclarecido (TCLE), e do “Manual de Recrutamento e Arrolamento”.

6. ARTIGOS

6.1. Artigo 1

Título: **"The impact of machine learning techniques in the study of bipolar disorder: a systematic review"**

Publicado na *Neuroscience & Biobehavioral Reviews*, Volume 80, páginas 538-554, em setembro de 2017. doi: 10.1016/j.neubiorev.2017.07.004.

Fator de impacto da revista: 8.299

Neuroscience and Biobehavioral Reviews 80 (2017) 538–554

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

The impact of machine learning techniques in the study of bipolar disorder: A systematic review

Diego Librenza-Garcia[a,b], Bruno Jaskulski Kotzian[b], Jessica Yang[c], Benson Mwangi[d], Bo Cao[d], Luiza Nunes Pereira Lima[b], Mariane Bagatin Bermudez[a,b], Manuela Vianna Boeira[b], Flávio Kapczinski[e], Ives Cavalcante Passos[b,f,*]

[a] Graduation Program in Psychiatry, Universidade Federal das Ciências da Saúde de Porto Alegre, Porto Alegre, RS, Brazil
[b] Bipolar Disorder Program, Laboratory of Molecular Psychiatry, Hospital de Clínicas de Porto Alegre (HCPA), Porto Alegre, RS, 90035-903, Brazil
[c] University of Texas at Austin College of Pharmacy, United States
[d] Department of Psychiatry and Behavioral Sciences, The University of Texas Science Center at Houston, Houston, TX, USA
[e] McMaster's Department of Psychiatry and Behavioral Neurosciences, Hamilton, Canada
[f] Graduation Program in Psychiatry and Department of Psychiatry, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, 90035-903, Brazil

ARTICLE INFO

Keywords:
Machine learning
Big data
Bipolar disorder
Suicide
Diagnosis
Support vector machine
Prediction
Predictive analysis
Pattern recognition
Neuroimaging

ABSTRACT

Machine learning techniques provide new methods to predict diagnosis and clinical outcomes at an individual level. We aim to review the existing literature on the use of machine learning techniques in the assessment of subjects with bipolar disorder. We systematically searched PubMed, Embase and Web of Science for articles published in any language up to January 2017. We found 757 abstracts and included 51 studies in our review. Most of the included studies used multiple levels of biological data to distinguish the diagnosis of bipolar disorder from other psychiatric disorders or healthy controls. We also found studies that assessed the prediction of clinical outcomes and studies using unsupervised machine learning to build more consistent clinical phenotypes of bipolar disorder. We concluded that given the clinical heterogeneity of samples of patients with BD, machine learning techniques may provide clinicians and researchers with important insights in fields such as diagnosis, personalized treatment and prognosis orientation.

1. Introduction

Bipolar disorder affects about 2% of the world's population with sub-threshold forms affecting an additional 2% of the population (Merikangas et al., 2007). According to the World Health Organization, bipolar disorder is among the 10 leading causes of disability-adjusted life years in young adults (Mathers et al., 2006). Rates of completed suicide in patients with bipolar disorder are 7.8% in men and 4.9% in women (Nordentoft, 2011), and life expectancy has been reported to decrease by 9 years in patients with bipolar disorder (Crump et al., 2013). Although several types of interventions may be used in order to prevent and treat mood episodes, they are frequently suboptimal, and about 60% of the patients relapse into depression or mania within two years of treatment initiation (Gitlin et al., 1995). In addition, current approaches to diagnosing bipolar disorder may not be completely effective, having an average delay of ten years between the first symptoms and the formal diagnosis (Lish et al., 1994). This framework illustrates important challenges in the current treatment approaches, diagnosis, and prevention in bipolar disorder.

Evidence-based medicine has helped us understand risk factors, optimal treatments and prognosis of bipolar disorder by using traditional statistical methods which primarily provide average group-level results (Sackett et al., 1996). However, in a recent article, Greenhalg and colleagues have called our attention to the fact that some statistically significant results may not represent a real benefit for an individual patient and that subjects in clinical trials may not always reflect the multimorbidity profile of real life patients (Greenhalgh et al., 2014). This may be particularly true in the field of bipolar disorder, where clinical heterogeneity is a very important factor. In light of these findings, techniques that aim at developing tailor-made psychiatric care to the individual, such as machine learning, have been gaining ground in psychiatric research (Huys et al., 2016).

Big data is a broad term used to denote volumes of large and complex measurements, as well as the velocity that data is created.

* Corresponding author at: Federal University of Rio Grande do Sul, Avenida Ramiro Barcelos, 2350, Zip Code: 90035-903, Porto Alegre, RS, Brazil.
E-mail addresses: diegolibrenzagarcia@gmail.com (D. Librenza-Garcia), brunokotzian@hotmail.com (B.J. Kotzian), jessica.yang10@gmail.com (J. Yang), benson.mwangi@gmail.com (B. Mwangi), cloudbocao@gmail.com (B. Cao), luiza.npl@gmail.com (L.N. Pereira Lima), mari.bermudez@yahoo.com.br (M.B. Bermudez), manuelaboeira@yahoo.co.uk (M.V. Boeira), flavio.kapczinski@gmail.com (F. Kapczinski), ivescp@yahoo.com.br (I.C. Passos).

http://dx.doi.org/10.1016/j.neubiorev.2017.07.004
Received 10 February 2017; Received in revised form 15 June 2017; Accepted 8 July 2017
Available online 18 July 2017

Another crucial characteristic of big data is the variety of levels in which data is created, from the molecular level, including genomics, proteomics and metabolomics, to clinical, sociodemographic, administrative, environmental, and even social media information (Passos et al., 2016b). In order to analyze big data, several machine learning methods (also known as pattern recognition techniques) have been developed during the last years. In short, one may say that first the algorithm analyses a 'training' dataset to establish a function able to distinguish individual subjects across groups. Once that has been done, the model can be applied to a new dataset, and the accuracy of the method can be measured in this new scenario. Later improvements of the model can be acquired, either by changing the algorithm or by performing additional feature reduction in the dataset (Lantz, 2015). As a result, these algorithms are ideal for assessing multifactorial disorders, and to estimate the probability of a particular outcome at an individual level (Mwangi et al., 2012).

The present study aims to review data in which bipolar disorder patients were assessed by using machine learning techniques regarding different outcomes. Our focus was mainly on studies that assessed diagnosis. However, we also included studies that assessed treatment, prognosis and development of data-driven phenotypes. We also provided a brief explanation about the most relevant principles of machine learning and its limitations in the supplementary material, since these techniques are relatively new in the field of psychiatry. Our overarching goal was to show how these new techniques are likely to support important clinical decisions in the forthcoming years.

2. Methods

2.1. Search strategy

We searched PubMed, Embase and Web of Science for articles published between January 1960, and January 2017 by using the following keywords: ("Big data" OR "Artificial Intelligence" OR "Machine Learning" OR "Gaussian process" OR "Cross-validation" OR "Cross validation" OR "Crossvalidation" OR "Regularized logistic" OR "Linear discriminant analysis" OR "LDA" OR "Random forest" OR "Naïve Bayes" OR "Least Absolute selection shrinkage operator" OR "elastic net" OR "LASSO" OR "RVM" OR "relevance vector machine" OR "pattern recognition" OR "Computational Intelligence" OR "Computational Intelligences" OR "Machine Intelligence" OR "Knowledge Representation" OR "Knowledge Representations" OR "support vector" OR "SVM" OR "Pattern classification") AND ("Bipolar Disorder" OR "Bipolar Disorders" OR "Manic-Depressive Psychosis" OR "Manic Depressive Psychosis" OR "Bipolar Affective Psychosis" OR "Manic-Depressive Psychoses" OR "Mania" OR " OR "Manic State" OR "Manic States" OR "Bipolar Depression" OR "Manic Disorder" OR "Manic Disorders" OR "Bipolar euthymic"). We also searched the reference lists to find potential articles to include. There were no language restrictions.

2.2. Eligibility criteria

This systematic review was performed according to the PRISMA statement (Liberati et al., 2009). Articles met the inclusion criteria if they assessed bipolar disorder patients using machine learning techniques. Technical and theoretical studies that used machine learning techniques but did not assess bipolar disorder patients were excluded. We also excluded studies that included only individuals below 18 years of age.

2.3. Data collection, extraction and statistical analysis

Two researchers (DLG and LNPL) independently screened titles and abstracts of the identified articles. They also obtained and read the full texts of potential articles, supervised by ICP who made the final decision in cases of disagreement. Data extracted from the articles included year of study publication, data used in the machine learning model (i.e., neuroimaging, blood biomarkers, clinical and demographical characteristics, among others), sample size, diagnoses assessed in the study, machine learning algorithm, and statistical measure of performance (i.e., accuracy, sensitivity, specificity, area under the curve, true positive, false positive, true negative and false negative). When this data was not available, we requested it from the authors.

We performed a meta-analysis of diagnostic accuracy with the classification studies. For this analysis, we included studies that used neuroimaging data (either structural or functional) to assess patients with bipolar disorder compared with healthy controls. Articles that assessed patients with bipolar disorder compared to patients with other psychiatric diagnosis were excluded. We also excluded studies performed in subjects with less than 15-year-old. We used the package *mada* from R (version number 3.3.1) to perform the analysis and to build the Summary Receiver Operating Characteristic (SROC) curve of sensitivity and specificity, as previously described (Reitsma et al., 2005).

3. Results and discussion

We found 757 potential abstracts and included 51 articles in the present review, being one of them added after reference screening (Fig. 1). A list of the included articles as well as its most relevant characteristics and findings is presented in Table 1 (classification studies) and tables S1 and S2 (clinical outcomes prediction and unsupervised learning studies). Most of the studies focused on diagnostic classification (38 studies) in order to distinguish bipolar disorder from schizophrenia, unipolar depression, healthy controls and other conditions. Of these, 11 used structural neuroimaging (Besga et al., 2012; Chen et al., 2014; Fung et al., 2015; Hajek et al., 2015; Koutsouleris et al., 2015; Mwangi et al., 2016; Redlich et al., 2014; Rocha-Rego et al., 2014; Sacchet et al., 2015; Schnack et al., 2014; Serpa et al., 2014), 13 used functional neuroimaging (Almeida et al., 2013; Anticevic et al., 2014; Arribas et al., 2010; Costafreda et al., 2011; Du et al., 2015; Frangou et al., 2017; Grotegerd et al., 2014, 2013; Jie et al., 2015; Kaufmann et al., 2017; Mourão-Miranda et al., 2012; Rive et al., 2016; Roberts et al., 2016), 5 used genetic analysis (Acikel et al., 2016; Chuang and Kuo, 2017; Dmitrzak-Weglarz et al., 2015; Pirooznia et al., 2012; Struyf et al., 2008), 4 used electroencephalographic measures (Erguzel et al., 2016, 2015; Johannesen et al., 2013; Khodayari-Rostamabad et al., 2010), 3 used neuropsychological tests, either alone or coupled with clinical observations and serum biomarkers (Akinci et al., 2013; Besga et al., 2015; Wu et al., 2016b), and 2 used a panel of serum biomarkers (Haenisch et al., 2016; Pinto et al., 2017). A total of 7 studies focused on predicting clinical outcomes, such as depression relapse and suicide (Salvini et al., 2015), mood changes (Faurholt-Jepsen et al., 2016; Gentili et al., 2017; Valenza et al., 2014, 2013) and suicide (Levey et al., 2016; Niculescu et al., 2015; Passos et al., 2016a). We found only 2 articles predicting treatment response or adverse effects (Castro et al., 2016; Wade et al., 2016), and 4 studies that used unsupervised or semi-supervised machine learning to identify homogeneous groups of patients (Bansal et al., 2012; Hall et al., 2012; Wahlund et al., 1998; Wu et al., 2016b).

3.1. Classification studies

Bipolar disorder particularly illustrates the dilemma of diagnostic systems solely based on clinical judgment, which may lead to misdiagnosis or treatment delay. It is known that bipolar disorder has an average delay of ten years between the first symptoms and formal diagnosis (Lish et al., 1994). It is also known that only 20% of patients with bipolar disorder who are experiencing a depressive episode are diagnosed with bipolar disorder within the first year of seeking treatment (Goldberg et al., 2001).

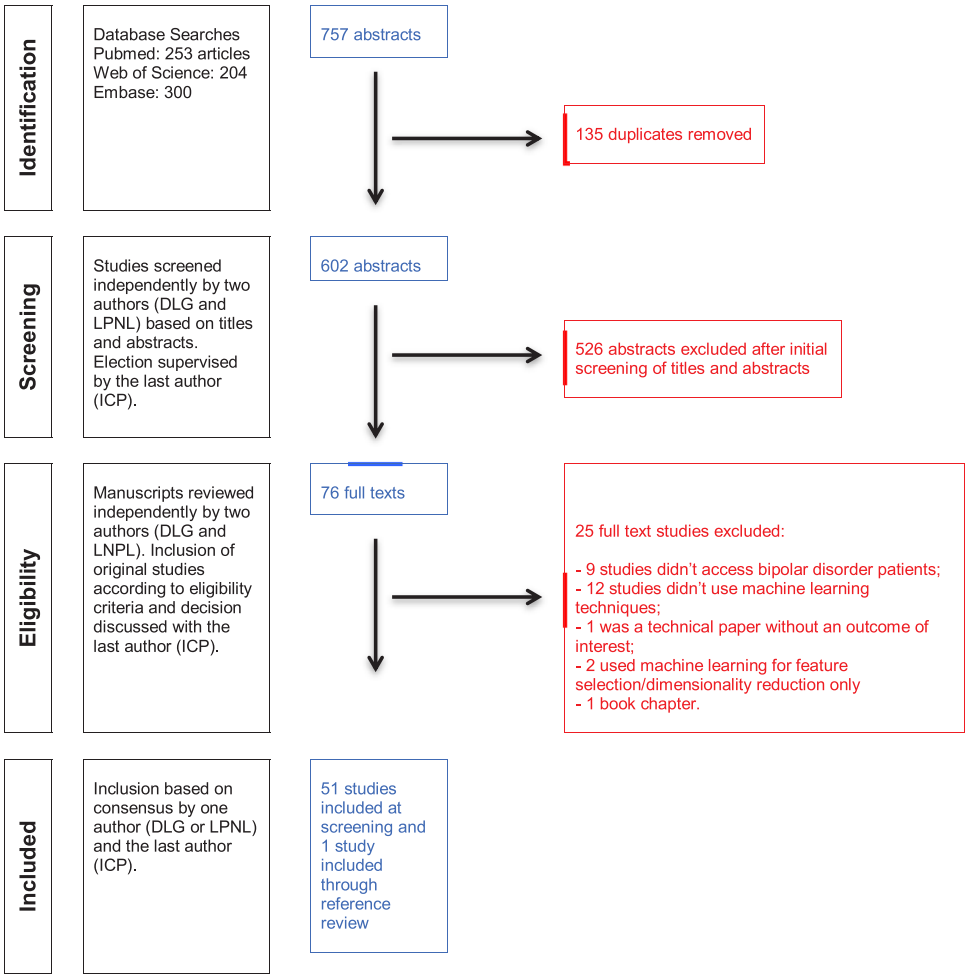


Fig. 1. Flowchart of review process and study selection.

Machine learning can aid clinicians by improving risk assessment and by allowing early detection of those at risk for bipolar disorder. In this systematic review, we included 24 studies that used structural or functional neuroimaging coupled with classification algorithms in order to distinguish between diagnostic classes or to predict diagnosis. For the meta-analysis of diagnostic accuracy, we included 12 studies based on our inclusion and exclusion criteria, comprising 16 different predictive models. We obtained AUCs of 0.698 for structural magnetic resonange imaging (sMRI), 0.754 for functional magnetic resonance imaging (fMRI) and 0.712 for both combined. Fig. 2 shows the Summary Receiver Operating Characteristic curve of sensitivity and specificity for classification studies that used neuroimaging data. Forest plots of the included studies are available in the supplementary material (Figs. S3-S5). Feature selection seems to play an important role on the final accuracy of these models, as neuroimaging studies tend to have few subjects but can produce a large number of features. The better is the selection of the relevant features, the better the model will perform (Erguzel et al., 2015; Guyon and Elisseeff, 2003; Mwangi et al., 2014; Sun et al., 2010; Tohka et al., 2016). Another 14 included studies used a variety of data, including clinical, sociodemographic, neuropsychological tests, electroencephalographic measures, among others, in the attempt to classify patients with bipolar disorder accurately.

3.1.1. Structural neuroimaging

We found 11 studies that used structural magnetic resonance imaging (MRI) in order to distinguish patients with bipolar disorder from patients with other psychiatric disorders or healthy controls (HC). In particular, many of these studies focused on differences between gray matter (GM) and white matter (WM). In a recent article, Mwangi et al. used GM and WM density in a large cohort of 256 patients to distinguish bipolar disorder from HC subjects, using the relevance vector machine (RVM) algorithm. Authors obtained accuracies of 70.3% for WM only, 64.9% for GM only, and 64% when both were combined (χ^2 $p < 0.005$). Another interesting finding was that early-stage bipolar disorder and HC were indistinguishable when RVM predicted probability scores were analyzed ($p = 0.05$), but intermediate-stage bipolar disorder and late-stage bipolar disorder differed significantly from HC ($p = 0.01$ and $p < 0.01$, respectively) (Mwangi et al., 2016). This supports previous evidence suggesting that bipolar disorder is a neuroprogressive disorder (Berk et al., 2011; Kapczinski et al., 2016). Another article used whole-brain GM from two independent samples to classify bipolar disorder, unipolar depression and HC subjects using support vector machine (SVM) and Gaussian process classifier (GPC) algorithms. Authors distinguished bipolar disorder from unipolar depression with accuracies of 75.9% (SVM, $p < 0.001$) and 79.3% (GPC, $p < 0.001$) in the first sample, and 65.5% (SVM, $p = 0.006$) and 65.5% (GPC, $p = 0.006$) in the second sample. When authors added WM to the model, similar to what happened in the study by Mwangi et al., there was no improvement in accuracy (Redlich et al., 2014). Schnack et al. also used GM density, but to discriminate between schizophrenia, bipolar disorder, and HC subjects. They trained an SVM model in one sample and used another sample to validate the model. Of note, the first sample was composed of bipolar disorder type I only patients, while the second sample had 24.2% of patients with bipolar disorder type II or not otherwise specified (NOS). The resulting accuracies were 55% and 63% of bipolar disorder and HC subjects correctly classified, and 66% and 65% of bipolar disorder and schizophrenia subjects correctly classified. The model was repeated excluding

Table 1
Machine learning studies classifying bipolar disorder and other diagnoses.

First author, year	Data utilized	Sample size and diagnosis[a]	Machine learning model	Accuracy	Other measures	Commentary
Classification studies using structural neuroimaging						
Besga et al. (2012)	sMRI and brain diffusion tensor imaging (BDI) in white matter.	57 subjects: - 20 with AD; - 12 with BD; - 25 HC.	SVM	100% (BD vs HC, BD vs AD)	Sensitivity: 100% (BD vs HC, BD vs AD) Specificity: 100% (BD vs HC, BD vs AD)	–
Chen et al. (2014)	3D sMRI Brain transformed into 2D images. SIFT algorithm to extract features from the brain images.	287 subjects: - 14 BD depressed patients; - 9 comparison subjects; - 48 mild AD; - 191 HC; - 27 PD.	SVM	57% (BD)	AUC 0.806	–
Fung et al. (2015)	MRI Brain comparisons on cerebral cortical thickness and surface area.	64 subjects: - 16 with BD; - 19 with UD; - 29 HC.	SVM	74.3%	Sensitivity: 62.5% Specificity: 84.2%	Effects of medication could not be estimated.
Hajek et al. (2015)	sMRI from affected and unaffected BD probands.	130 subjects: - 45 unaffected subjects; - 36 affected relatives of BD patients; - 49 HC.	SVM GPC	SVM: 68.9% (Un HR vs. HC); 59.7% (Af HR vs. HC). GPC: 65.6% (Un HR vs. HC)	Sensitivity: SVM: 75.6% (Un HR vs. HC); GPC: 71.1% (Un HR vs. HC); SVM: 58.3% (Af HR vs. HC). Specificity: SVM: 62.2% (Un HR vs. HC); GPC: 60% (Un HR vs. HC); SVM: 61.1% (Af HR vs. HC).	More left-handed participants in the unaffected HR group.
Koutsouleris et al. (2015)	MRI-based multivariate pattern analysis.	846 subjects: - 158 with SCZ; - 104 with MDD; - 35 with BD; - 23 with FEP; - 89 with ARMS; - 437 HC.	PCA (feature selection) SVM (model)	74% of BD subjects classified as MDD; the rest classified as SCZ	Sensitivity: 79.8% Specificity: 72.2%	–
Mwangi et al. (2016) *predictive post-hoc.	sMRI of gray and white matter.	256 subjects: - 128 with BD; - 128 HC.	RVM	70.3% (RVM/WM); 64.9% (RVM/GM); 64% (RVM/WM + GM).	Sensitivity: 66.4% (RVM/WM); 58.6% (RVM/GM); 59% (RVM/WM + GM). Specificity: 74.2% (RVM/WM); 71.1% (RVM/GM); 70% (RVM/WM + GM). AUC: 0.72 (RVM/WM); 0.7 (RVM/GM).	–

(continued on next page)

Table 1 *(continued)*

First author, year	Data utilized	Sample size and diagnosis[a]	Machine learning model	Accuracy	Other measures	Commentary
Redlich et al. (2014)	sMRI of gray matter. The features were compared using Voxel-based morphometry.	174 subjects: - 58 with BD; - 58 with UD; - 58 HC.	SVM GPC	SVM: 65.5%–75.9% (UD vs. BD); GPC: 65.5%–79.3% (UD vs. BD).	–	Medication load was higher for individuals with BD than for those with UD.
Rocha-Rego et al. (2014)	sMRI of gray and white matter.	80 subjects: - 40 with BD type I; - 40 healthy controls.	GPC	69–78%	Sensitivity: 64–77% Specificity: 69–99%	–
Sacchet et al. (2015)	MRI subcortical gray matter volumes.	193 subjects - 40 with BD; - 57 with MDD; - 35 with past but not current MDD; - 61 HC.	SVM RFE	MDD vs BD: 59.45% (with RFE)	–	Some of the participants in the MDD group may convert to BD.
Schnack et al. (2014)	sMRI, gray matter density images.	198 subjects: - 66 with SCZ; - 66 with BD; - 66 HC.	SVM	BD vs. SCZ: 66%/65%; BD vs. HC: 55%/65%.	Sensitivity: BD vs. SCZ: 79%; BD vs. HC: 70%. Specificity: BD vs. SCZ: 48%; BD vs. HC: 49%.	All SCZ patients have received antipsychotic medications.
Serpa et al. (2014)	Gray and white matter and Regional Analysis of Volumes Examined in Normalized Space (RAVENS) analyzed in sMRI.	117 subjects: - 27 with BD-I; - 19 with psychotic MDD; - 71 HC.	SVM	BD-I vs. HC: 66.1%; BD-I vs. psychotic MDD: 54.76%.	Sensitivity: BD-I vs. HC: 39.1%; BD-I vs. psychotic MDD: 57.9%. Specificity: BD-I vs. HC: 84.8%; BD-I vs. psychotic MDD: 52.1%.	A significant proportion of patients were using antipsychotic, which is associated with both gray matter and white matter reductions.
Classification studies using functional neuroimaging						
Almeida et al. (2013)	Arterial Spin Labelling (ASSL) measuring blood flow at rest of anterior cingulate cortex (ACC).	54 subjects: - 18 with BD; - 18 with UD; - 18 HC.	SVM	81%	Sensitivity: 83% Specificity: 78%.	All subjects were female; results presented here are for the subgenual ACC model (BD vs. UD).
Anticevic et al. (2014)	Resting-State fMRI of thalamic coupling via individual-specific anatomically derived thalamic seeds.	294 subjects: - 90 with SCZ; - 67 with BD; - 137 HC.	SVM	61.7%	–	–
Arribas et al. (2010)	fMRI collected while subjects are performing two runs of an auditory oddball (AOD) task.	95 subjects: - 21 SCZ;	NN PPMS	NA	AUCs: 0.807–0.820 (HC vs. non-HC)	–

(continued on next page)

Table 1 *(continued)*

First author, year	Data utilized	Sample size and diagnosis[a]	Machine learning model	Accuracy	Other measures	Commentary
		- 14 BD; - 25 HC.				
			AIC MDL		0.878–0.890 (BD vs. non-BD) 0.885–0.902 (SCZ vs. non-SCZ)	
Costafreda et al. (2011)	fMRI of neural responses to verbal fluency task.	104 subjects: - 32 with SCZ; - 32 with BD; - 40 HC.	SVM	79% (BD)	Sensitivity: 56% Specificity: 89%	Patients were receiving medication.
Du et al. (2015)	Resting state brains fMRI networks.	93 subjects: - 20 with SCZ; - 20 with BD; - 20 with SA in manic episode; - 13 with SA in depressive episode; - 20 HC.	MSVM-RFE t-SNE	68.75% (overall)	–	Model misclassified 2 out of 20 BD subjects.
Frangou et al. (2017)	Task-based functional magnetic resonance during the n-back working memory task.	120 subjects: - 30 BD type I patients; - 30 MDD (first-degree relatives of the included BD patients); - 30 HC (unrelated with the included BD patients); - 30 HC (first-degree relatives of the included BD patients).	GPC	83.5% (BD vs unrelated HC) 76.9% (BD vs. MDD relatives) 81.8% (BD vs. related HC)	BD vs unrelated HC: Sensitivity: 84.6% Specificity: 92.3% PPV: 91% NPV: 85%	Analysis performed comparing 0-back, 1-back, 2-back and 3-back contrast; most comparisons were not statistically significant (not shown here).
Grotegerd et al. (2013)	fMRI of amygdala excitability to emotional faces.	30 subjects: - 10 with BD; - 10 with UD; - 10 HC.	SVM GPC.	SVM: 80% (BD vs. HC). GPC: 70% (BD vs. HC).	Sensitivity: 90% (Negative vs. Neutral). Specificity: 60% (Negative vs. Neutral).	–
Grotegerd et al. (2014)	Brain fMRI of amygdala responsiveness to negative > neutral stimuli.	66 subjects: - 22 with BD; - 22 with MDD; - 22 HC.	SVM GPC.	79.6% (sad > happy contrast)	–	–
Jie et al. (2015)	sMRI and resting-state fMRI.	69 subjects: - 21 with BD; - 25 with MDD; - 23 HC.	SVM-FoBa	92.07% (BD vs. MDD); 80.78% (BD vs. HC).	–	–
Kaufmann et al. (2017)	fMRI acquired while performing five different cognitive tasks.	323 subjects: - 136 HC; - 97 BD; - 90 SCZ.	rLDA	BD classification: 91.74% (2-back)	–	Groups differ in gender (BD had more females).

(continued on next page)

Table 1 *(continued)*

First author, year	Data utilized	Sample size and diagnosis[a]	Machine learning model	Accuracy	Other measures	Commentary
				69.13% (0-back) 63.93% (GNG) 52.13% (FacePos) 49.55% (FaceNeg)		
Mourão-Miranda et al. (2012)	Emotional face tasks during fMRI.	54 subjects: - 18 with BD; - 18 with UD; - 18 HC.	GPC	61% (BD – intense happy versus neutral faces)	Sensitivity: 72% Specificity: 50%	–
Rive et al. (2016)	Used sMRI and RS-fMRI data of areas implicated in mood disorders.	91 subjects: - 10 depressed BD patients; - 26 remitted BD patients; - 22 depressed MDD patients; - 23 remitted MDD (22 for the RS-fMRI due to missing data).	GPC, SVM	Best accuracy for default mode network MDDd vs BDd (SVM): 80.5%	Default mode network MDDd vs BDd (SVM): Sensitivity: 70.0% Specificity: 90.9% PPV: 77.8% NPV: 87.0%	Included only medication-free subjects; included both depressed and remitted patients; depressed group with more previous depressive episodes.
Roberts et al. (2016)	Resting-state fMRI connectivity data of the left inferior frontal gyrus (IFG).	200 subjects: - 49 BD patients; - 71 AR subjects; - 80 HC.	Multiclass SVM	Three group classifier: HC: 58.0% AR: 64.5% BD: 70.5% Overall (mean): 64.3%	Sensitivity HC: 56.3% AR: 46.5% BD: 30.6% Specificity HC: 59.2% AR: 74.4% BD: 83.4%	Included only young BD subjects (16–30 years); AR subjects had at least one first-degree relative with BD; chance rate 40.8% due to unequal group size.
Classification studies using genetic analysis						
Acikel et al. (2016)	Whole-genome genotyping; 693 candidate SNPs that remained after cleaning and filtering.	2371 subjects: - 604 BD; - 1767 HC.	RF NB k-NN MDR	0.734 (RF) 0.702 (NB) 0.733 (kNN) 0.647 (two-way MDR) 0.721 (three-way MDR)	Sensitivity 0.998 (RF) 0.734 (NB) 0.954 (kNN) 0.664 (two-way MDR) 0.883 (three-way MDR) PPV 0.743 (RF) 0.845 (NB) 0.754 (kNN) 0.675 (two-way MDR) 0.772 (three-way MDR)	Data of patients with bipolar-related disorders were excluded.

(continued on next page)

Table 1 *(continued)*

First author, year	Data utilized	Sample size and diagnosis[a]	Machine learning model	Accuracy	Other measures	Commentary
Chuang and Kuo (2017)	Two individual genome-wide association datasets comprising an initial number of 1,992,730 SNPs.	4488 subjects: - 1956 BD; - 2532 HC.	RF	NA	AUCs (testing dataset) 0.944 (0.935–0.953) for all 289 markers 0.924 (0.913–0.935) for 121 optimal markers AUCs (validation dataset) 0.702 (0.681–0.723) for all 289 markers 0.639 (0.617-0.662) for 121 optimal markers	Only Caucasian subjects.
Dmitrzak-Weglarz et al. (2014)	42 SNPs from four genes of interest.	1379 subjects: - 229 with UD; - 515 with BD; - 635 HC.	CART	61% (HC vs. non-HC)	Sensitivity: 71% Specificity: 50%	Only ten of the analyzed polymorphism have predictive value.
Pirooznia et al. (2012)	GWAS and SNP datasets.	3625 subjects: - 2191 with BD; - 1434 HC.	BN SVM RF LR RBFN	NA	AUCs ranging from 0.482–0.556 in all algorithms	–
Struyf et al. (2008)	Gene expression, demographic and clinical data.	332 subjects: - 115 with SCZ; - 105 with BD; - 112 HC.	SVM NSC DT EoV NB k-NN	–	AUC 0.92 (genetic expression only) and 0.97 (genetics plus clinical and demographical data) using SVM	–
Classification studies using electroencephalografic measures						
Erguzel et al. (2015)	Electroencephalography (EEG) biomarker. Values calculated from Alpha, Theta and Delta frequency bands.	101 subjects: - 46 with BD in depressive episode; - 55 with UD.	SVM	SVM: 62,37%; PSO-SVM: 73,26%; GA-SVM: 75,24%; ACO-SVM: 78,21%; IACO-SVM: 80,19%.	AUCs: SVM: 0.631; PSO-SVM: 0.739; GA-SVM: 0.776; ACO-SVM: 0.779; IACO-SVM: 0.793.	No control-group.
Erguzel et al. (2016)	Cordance (quantitative electroencephalographic method); alpha and theta frequency bands.	89 subjects - 31 with BD; - 58 with UD.	PSO for feature selection and ANN for training	73.03 (ANN only) 83.87 (PSO-ANN)	ANN AUC: 0.757 Sensitivity: 64.52% (BD) PSO-ANN AUC: 0.905 Sensitivity: 83.87% (BD)	–

(continued on next page)

Table 1 *(continued)*

First author, year	Data utilized	Sample size and diagnosis[a]	Machine learning model	Accuracy	Other measures	Commentary
Johannesen et al. (2012)	Neurophysiological Endophenotypes (P50 and P300).	150 subjects: - 50 with SCZ; - 50 with BD; - 50 HC.	MLR	72% (BD vs. SCZ)	Sensitivity: 74% Specificity: 70%	Groups may differ in P200 and N200 ERPs.
Khodayari-Rostamabad et al. (2010)	EEG data with feature selection.	207 subjects: - 64 with MDD; - 40 with SCZ; - 12 with BD; - 91 HC.	MFA	92.7%	–	–
Classification studies using neuropsychological tests						
Akinci et al. (2013)	A video-based eye pupil detection system for diagnosing bipolar disorder.	85 subjects: - 40 with BD; - 55 HC.	SVM	96.36%	–	–
Besga et al. (2015)	Clinical observations, neuropsychological tests and specific blood plasma biomarkers.	95 subjects: - 37 with AD; - 32 with late-onset BD; - 26 HC.	SVM RF CART LASSO	95.76% (HC vs. LOBD); 90.26% (AD vs. LOBD).	–	CART was the best classifier.
Wu et al. (2016a)	CANTAB.	42 subjects: - 21 with BD; - 21 HC.		71%	Sensitivity: 76% Specificity: 67%	Only considered euthymic BD patients.
Classification studies using serum biomarkers						
Haenisch et al. (2016)	Proteomics to evaluate blood biomarkers of BD.	907 subjects: - 249 with BD; - 122 with pre-diagnostic BD; - 75 with pre-diagnostic SCZ; - 90 first onset MDD; - 371 HC.	LASSO	NA	AUC: 0.79 (BD vs HC) 0.91 (BD vs SCZ)	–
Pinto et al. (2017)	Peripheral biomarkers: BDNF, IL-6, IL-10, CCL11, gluthatione S-transferase, glutathione peroxidase.	60 subjects: - 20 HC; - 20 with BD; - 20 with SCZ.	SVM	72.5% (BD vs. HC) 77.5% (SCZ vs. HC)	BD vs. HC: Specificity = 73.68% Sensitivity = 71.42% PPV = 75% NPV = 70% BD vs. SCZ: Specificity = 78.94% Sensitivity = 76.19% PPV = 80% NPV = 75%	Euthymic BD patients only; SCZ patients without acute psychosis; subjects matched by age and gender.

Abbreviations;:*AD,* Alzheimer disease; AIC, Akaike's information criterion; ANNs, Artificial Neural Networks; AR, at-risk subjects; ARMS, at-risk mental state for psychosis; AUC, Area under the curve; BD, bipolar disorder; BDd, bipolar disorder depressed; BDr, bipolar depression remitted; BDNF, brain-derived neurotrophic factor; BN, Bayesian networks; CANTAB, Cambridge Neurocognitive Test Automated Battery; CART, Classification and Regression trees; CCL11, eotaxin-1; DSM,

Diagnostic and Statistical Manual of Mental Disorders; DT, Decision trees; ECT, electroconvulsive therapy; EoV, Ensemble of voters; FaceNeg, negative facial stimuli recognition task; FacePos, positive facial stimuli recognition task; GNG, go/no-go task; GPC, Gaussian Process Classifiers; GWAS, Genome-wide association study; HC, healthy controls; IL-6, interleukin-6; IL-10, interleukin-10; k-NN, k-Nearest neighbor; LASSO, Least Absolute Shrinkage and Selectin Operator; LR, Logistic regression; MDD, major depressive disorder; MDDd, Major depressive disorder depressed; MDD_r, major depressive disorder remitted; MDL, Minimum Description Length; MDR, Multifactor dimensionality reduction; MFA, Mixture Factor Analysis; MLR, Multivariate logistic regressions; MSVM-RFE, Multiclass support vector machine recursive feature elimination; NB, Naïve Bayes; NN, Neural Networks; NPV, Negative predictive value; NSC, Nearest shrunken centroids; PCA, Principal component analysis; PD, Parkinson disorder; PPMS, Posterior Probability Model Selection; PPV, Positive predictive value; PSO, particle swarm optimization; RBFN, Radial basis function network; RF, Random Forest; RFE, Recursive feature elimination; rLDA, regularized linear discriminant analysis; RS-fMRI, resting state functional magnetic resonance imaging; RVM, Relevance Vector Machine; SA, schizoaffective disorder; SCZ, schizophrenia; SIFT, Scale Invariant Feature Transform; SNP, Single-nucleotide polymorphisms; SVM, Support Vector Machine; SVM-FoBa, Support Vector Machine with Forward-Backward Search Strategy; t-SNE, t-distributed stochastic neighbor embedding; UD, unipolar disorder.

[a] All studies used DSM-IV criteria for diagnosis, except when specified otherwise. Akinci et al., (2013), Arribas et al. (2010), and Redlich et al., (2014) didn't specify diagnostic criteria.

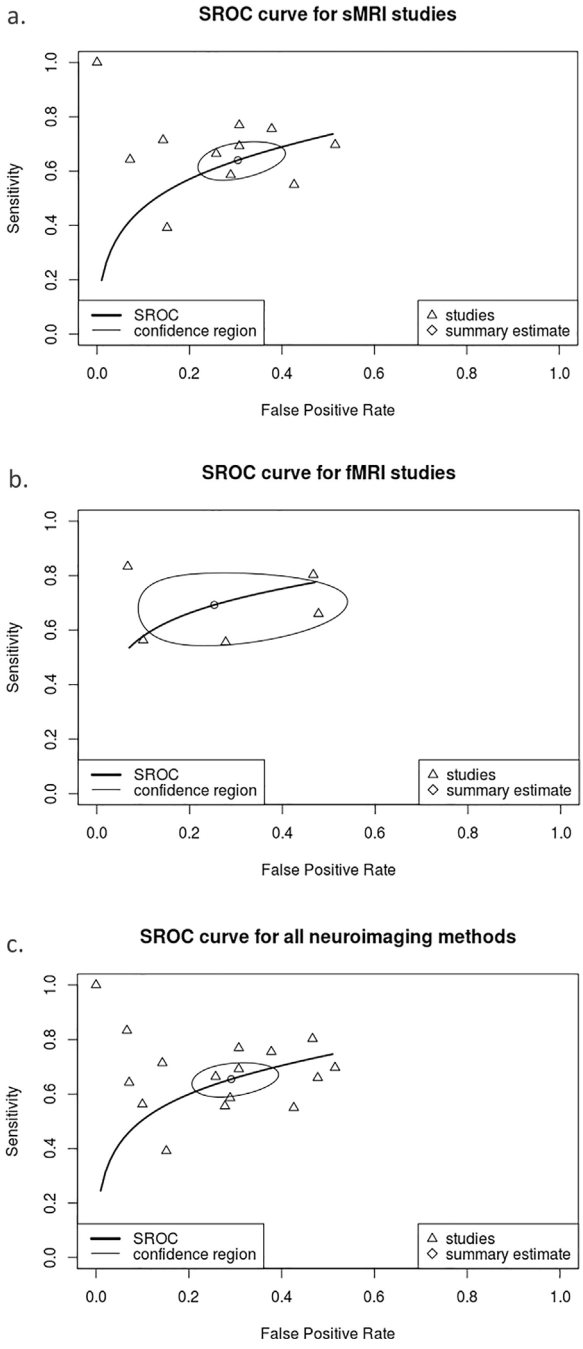


Fig. 2. Summary Receiver Operating Characteristic (SROC) curve of sensitivity and specificity for classification studies using neuroimaging (bipolar disorder versus healthy controls). (a) Satructural neuroimaging studies. (b) Functional neuroimaging studies. (c) Structural and functional neuroimaging studies.

individuals on lithium therapy and showed no significant difference in the resulting accuracy (Schnack et al., 2014). In another study, GM volumes of selected brain regions were used in order to distinguish between major depressive disorder (MDD), euthymic bipolar disorder type I, remitted MDD and healthy controls. The bipolar disorder versus MDD classification had an accuracy of 59.45% while using a linear SVM with Recursive Feature Elimination (RFE). However, the other results were not statistically significant (Sacchet et al., 2015).

Fung et al. used cortical thickness and surface area, morphometric features with distinct genetic determinants, to distinguish between bipolar disorder, MDD and HC subjects. Using an SVM algorithm, they obtained an accuracy of 74.3% ($p = 0.028$), with 62.50% sensitivity

and 84.20% specificity (Fung et al., 2015). In another study, the authors used feature-based morphometry combined with an SVM algorithm to select regions of interest prior to classification and compared results to voxel-based morphometry (VBM) and deformation-based morphometry (DBM), two already established methods. The new method was applied to identify bipolar disorder patients in the corresponding sample, resulting in an accuracy of 57% (Chen et al., 2014). Koutsouleris et al. aimed to classify individuals with schizophrenia, bipolar disorder, MDD and at-risk mental states (ARMS). The authors hypothesized that their neuroanatomical signatures varied on a continuum between schizophrenia and MDD. They trained an SVM machine learning model using a sample of schizophrenia and MDD patients, and then applied the model to a sample containing bipolar disorder patients. The algorithm classified 74% of bipolar disorder as MDD − suggesting proximity of these disorders in this continuum. Individuals with first-episode psychosis, ultra-high risk and low-risk for psychosis were mostly classified in the schizophrenia group, suggesting these conditions are more related to this extreme end of the continuum (Koutsouleris et al., 2015). Another study used SVM-based pattern classifier to distinguish between first-episode psychotic mania (bipolar disorder type I), first-episode psychotic MDD and HC. The authors achieved accuracies of 66.1% when differentiating bipolar disorder type I versus HC and 54.76% when distinguishing bipolar disorder type I from MDD. The low accuracies obtained in this scenario limits its clinical application, but unlike other studies in which subjects had greater illness duration, this study tried to create a useful tool to detect early stages of the disease (Serpa et al., 2014). Finally, one study used diffusion tensor imaging (DTI) features coupled with WM integrity analysis to distinguish between late-onset bipolar disorder, Alzheimer's Disease (AD) and HC. Using an SVM algorithm trained and tested on fractional anisotropy (FA), the authors achieved 100% accuracy in all discriminations tasks (Besga et al., 2012).

Another study examined the feasibility of using pattern recognition techniques to distinguish patients with bipolar disorder from HC in two independent study populations (Rocha-Rego et al., 2014). The samples consisted of two cohorts of remitted bipolar disorder type I patients. The authors applied a GPC algorithm to GM and WM structural magnetic resonance imaging data. The diagnostic accuracy of the GPC for GM was 73% in study population 1 and 72% in study population 2; the sensitivity and specificity of the classification were respectively 69% and 77% in study population 1 and 64% and 99% in study population 2. The diagnostic accuracy of the GPC for WM was 69% in study population 1 and 78% in study population 2; the sensitivity and specificity of the WM classification were both 69% in study population 1 and 71% and 86% in study population 2, respectively. Nevertheless, the two cohorts differed in medication status, IQ, age and age of disease onset. A different study used both SVM and GPC to identify changes in brain structures using sMRI in unaffected and affected bipolar offspring. Authors obtained 68.9% accuracy in distinguishing high-risk unaffected subjects from controls and 59.7% accuracy in differentiating affected offspring from controls. Results showed that WM from the inferior and middle frontal gyrus, as well as from the temporal gyrus and the precuneus contributed the most to distinguish high genetic risk subjects from low-risk ones. These results, however, were not repeated when analyzing GM (Hajek et al., 2015).

3.1.2. Functional neuroimaging

Functional neuroimaging was also employed in some studies. Grotegerd et al. used SVM and GPC coupled with functional neuroimaging to distinguish depressed bipolar from unipolar patients. Subjects underwent neuroimaging acquisition while performing passive-viewing tasks on different facial expressions. Authors obtained accuracies of 90% (happy versus neutral face), 75% (negative versus neutral faces) and 80% (combining the previous ones) with the SVM model. GPC, however, did not perform so well (respectively 70, 70 and 75% accuracy) (Grotegerd et al., 2013). In another study of the same group, patients with bipolar disorder and MDD were presented with a rapid stimulus of either a happy, sad, or neutral face, then a neutral-face mask, and asked to describe the emotion of the flashed face. The fMRI activity in the amygdala obtained while identifying happy versus sad faces was the most predictive in differentiating bipolar disorder versus MDD, with an accuracy of 75% using SVM and 79.6% using GPC, both statistically significant ($p = 0.008$ and 0.002, respectively). It is important to state that the samples from these two studies did not overlap (Grotegerd et al., 2014). Another study used fMRI to determine differences in brain activity in patients with bipolar disorder, MDD and HC while they were presented with intense happy, mild happy and neutral faces and were asked to distinguish them. They trained a GPC algorithm using whole-brain activity patterns and were able to distinguish bipolar disorder from unipolar depression using mild happy faces task, obtaining 67% accuracy with 72% specificity and 61% sensitivity ($p = 0.018$). However, this result did not remain statistically significant after correcting for multiple comparisons (Mourão-Miranda et al., 2012).

Frangou and colleagues used fMRI acquisition during the n-back working memory task to differentiate between patients with bipolar disorder from unrelated healthy controls, and their first-degree relatives that were either healthy individuals or had a MDD diagnosis. The authors achieved accuracies of 83.5 (bipolar disorder vs. unrelated HC), 73.1 (bipolar disorder vs. MDD relatives) and 81.8% (bipolar disorder vs. related HC) using the data from the 3-back vs. 0-back contrast. Most of the comparisons between the other contrast data, however, were not statistically significant (Frangou et al., 2017). Kaufmann and colleagues used functional network connectivity while participants performed five different cognitive tasks during fMRI scans in bipolar disorder ($n = 97$), schizophrenia ($n = 90$) and HC (136) subjects to identify these tasks within each diagnosis class, aiming to correlate brain structure and brain function. Using a regularized linear discriminant analysis to create a 5-class classifier, authors obtained an accuracy of 91.74% within bipolar disorder class to classify the 2-back test based on the connectivity profile of fMRI data. Accuracies for classifying the other four tasks ranged from 49.55-69.13%. Similar classification accuracies were obtained for the three subjects groups, and three of the five tasks could be classified with accuracies over 60% in all of them (Kaufmann et al., 2017).

In another study, authors used blood-oxygen-level dependent (BOLD) signals within the default-mode network and temporal lobe measured by fMRI to differentiate between bipolar disorder, schizophrenia and HC subjects while they performed two runs of an auditory oddball (AOD) tasks. After dimensionality reduction of the fMRI data, four different machine learning algorithms were applied to distinguish bipolar disorder from non-bipolar disorder subjects, with area under the curves (AUCs) ranging from 0.878-0.890. The three-way correct classification rate (CCR) of the algorithms was between 70.1 and 70.9% (Arribas et al., 2010). Costafreda et al. used fMRI to determine neural responses in bipolar disorder, schizophrenia and healthy control subjects while they performed a phonological verbal fluency task. The authors were able to correctly identify patterns in neuronal responses in patients with bipolar disorder with an accuracy of 79%, sensitivity of 56% and 89% specificity ($p < 0.001$). However, the prediction accuracy for patients with bipolar disorder was much lower than that for patients with schizophrenia, which reached 92% accuracy (Costafreda et al., 2011). Du et al. used resting state brain networks to build a model to differentiate between bipolar disorder, schizophrenia, HC, and schizoaffective disorder with manic episodes (SADM) or with depressive episodes only (SADD). 93 subjects were used as the testing dataset, and the model was validated in a new sample consisting of 16 subjects. Using SVM to both the feature selection and the pattern classification tasks, authors obtained a 68.75% overall accuracy in the validation sample(Du et al., 2015). Another study used thalamus seed-based connectivity analyses in order to distinguish between schizophrenia, bipolar disorder and HC. The authors hypothesized that patients with

bipolar disorder also experience thalamo-cortical communication disturbances based on similar previous findings in patients with schizophrenia. Using a linear SVM with leave-one-out cross-validation, the authors distinguished bipolar disorder from HC with an accuracy of 61.7% ($P < 0.038$) (Anticevic et al., 2014).

Jie et al. combined anatomical and functional data in order to distinguish bipolar disorder patients from MDD and HC in three public datasets. The authors first analyzed fractional amplitude of low-frequency fluctuation (fALFF), obtained through fMRI, and the voxel-based morphometry of voxel-wise gray matter (VBM-GM) volume, obtaining classification accuracies for each method. After that, they combined both sets of data in a multimodal analysis in order to improve accuracy. They obtained accuracies of 92.07% when discriminating bipolar disorder from MDD and 80.78% when discriminating bipolar disorder from HC, higher than those for fALFF or VBM-GM alone. Prior to the multimodal analysis, the authors also developed a new algorithm called SVM with forward-backward search strategy (SVM-FoBa) and tested it against four already established methods of feature selection in three public datasets (Jie et al., 2015).

Rive et al. used both sMRI (gray matter volume) and resting-state fMRI to distinguish depressed and remitted patients with bipolar disorder or MDD using GPC and SVM models. Both GPC and SVM had poor performances to differentiate remitted patients, although default mode network (fMRI) was able to distinguish depressed patients with 69.1% (GPC, $p = 0.02$) and 80.5% (SVM, $p = 0.01$). Results interpretation are limited because the study had a small sample and included MDD, bipolar disorder type I, and bipolar disorder type II in the depressed group. Also, the depressed group had statistically significant more depressive episodes than the remitted group (Rive et al., 2016). Another study used data from resting functional connectivity of the left inferior frontal gyrus to create a SVM model capable of differentiate bipolar disorder patients, at-risk subjects (defined as having at least one first-degree relative with bipolar disorder), and healthy controls. Unlike most of the studies, authors used a multi-class classifier, i.e., the model was used to distinguish the three classes at once, instead of in a pairwise fashion. They obtained an overall accuracy of 64.3%, with individual accuracies of 58% (HC), 64.5% (AR) and 70.5% (bipolar disorder). Of note, chance rate was 40.8% because of different size groups (Roberts et al., 2016).

One article used arterial spin labelling (ASL) to analyze anterior cingulate cortex (ACC) blood flow in order to distinguish between bipolar disorder, unipolar depression and HC. ASL is a non-invasive perfusion magnetic resonance that quantifies cerebral blood flow (Petcharunpaisan et al., 2010). Pattern recognition analysis with SVM of the subgenual ACC subdivision differentiated bipolar disorder from unipolar depression with 81% accuracy. No region was significant to discriminate bipolar disorder from HC. However, it is important to note that this study used only female subjects and had a small sample size (18 patients in each group) (Almeida et al., 2013).

One critical issue while using neuroimaging is generalization. Most of the studies had small samples and did not repeat the analysis in unseen samples, which can limit the obtained results (Du et al., 2015). Also, the main application of the classifiers using machine learning coupled with neuroimaging may not be to detect already established diagnosis, which remains a clinical decision, unless the diagnosis would be data-driven. Instead, the great impact of this combined approach would be to predict which individuals are at risk of progression to a major psychiatric disorder, allowing early and precise interventions.

3.1.3. Genetic analysis

Five studies used machine learning algorithms coupled with genetics in order to build models to differentiate patients with bipolar disorder from patients with other disorders. Struyf et al. combined gene expression data from the Stanley Neuropathology Consortium with demographic and clinical data to build six algorithms for classifying bipolar disorder and schizophrenia patients. Using genetic data only, the SVM model was able to distinguish bipolar disorder from controls with an AUC of 0.92. However, after combining genetic data with demographic and clinical data, the performance was improved to an AUC of 0.97. The SVM model outperformed all other algorithms, all of which obtained AUCs ranging from 0.60–0.73 (genetic only) and 0.60–0.90 (genetics plus clinical and demographic data). These results illustrate the application and effectiveness of SVM in large datasets, as well as the superiority of multimodal signatures using data from different biological and clinical levels (Struyf et al., 2008). Pirooznia et al. compared five machine learning algorithms to an established polygenic scoring approach to classify patients with bipolar disorder and schizophrenia. The model was built using the Bipolar Genome Study dataset and tested using the Wellcome Trust Case Control Consortium dataset, both containing large samples with thousands of single-nucleotide polymorphisms (SNP) available. The analysis was conducted in four parts: two whole-genomes with 3514 SNPs (WG1) and 14364 SNPs (WG2) and two brain-expressed SNPs with 1252 SNPs (BE1) and 5366 SNPs (BE2). In WG1, the Bayesian networks (BN) algorithm was superior to the polygenic scoring approach, with an AUC of 0.55 against 0.549. The other four algorithms performed worse, with AUCs ranging from 0.482 to 0.545. In WG2, the polygenic scoring approach outperformed all machine learning algorithms. Results were similar with BE1 and BE2 analysis, with BN slightly superior to the polygenic scoring approach (Pirooznia et al., 2012).

One study used gene variants from four core period proteins involved in circadian rhythm to differentiate unipolar depression, bipolar disorder, and HC. Using a Classification and Regression Trees (CART) algorithm, the authors classified unipolar depression from HC with an accuracy of 61%. However, the authors did not find any significant model for classifying bipolar disorder against unipolar depression and HC (Dmitrzak-Weglarz et al., 2015). Acikel et al. used three data mining methods (Random Forest, Naïve Bayes and k-Nearest Neighbors) and multifactorial dimensionality reduction (MDR) in data obtained from the Whole-Genome Association Study of Bipolar Disorder with the purpose of finding SNP associations implicated in bipolar disorder. After cleaning and filtering the SNPs, the remaining 693 SNPs were used to build four models, with accuracies of 73.4% (RF; 15 SNPs), 70.2% (NB; 13 SNPs), 73.3% (kNN; 10 SNPs), 64.7% (two-way MDR; 6 SNPs) and 72.1% (three-way MDR; 6 SNPs), based respectively in 16, 13, 10, 6 and 6 SNPs. Of note, while all the three data mining methods found the same top six SNPs, MDR found different SNPs, probably due to its ability to find associations between different SNPs (Acikel et al., 2016).

Another study aimed to determine a genetic risk model in a Caucasian sample using two genome-wide association datasets to identify bipolar disorder and HC subjects, one for model construction and another one for validation. They started with 1,992,730 intersection SNPs between the datasets, and obtained an AUC of 0.944 (0.935–0.953) in the training dataset and 0.702 (0.681–0.723) in the validation dataset while using the 289 selected candidate markers, and AUCs of 0.924 (0.913–0.935) and 0.639 (0.617–0.662) in training and validation sets, respectively, with the 121 identified optimal markers. Authors also created a risk model multiplying the numbers of risk allele by the beta regression coefficient of each marker, with poor performance in the validation step (Chuang and Kuo, 2017). While the results are promising in comparison to genetic risk models previously proposed for major psychiatric disorders, the model is still not feasible for clinical use, with significant lower discrimination ability in the external datasets. Reasons for that may be the small sample size for a genome-wide association study and the fact that bipolar disorder is a complex disease in which environmental factors play a significant role, none of which used in the present model.

3.1.4. EEG features

Four studies used electroencephalogram (EEG) biomarkers to assess bipolar disorder diagnosis. Erguzel et al. performed two studies with

different feature selection algorithms prior to the training model. In the first, authors used EEG coherence, a quantitative EEG biomarker that has the potential to reflect differing brain dynamics between bipolar disorder and MDD patients, using Improved Ant Colony Optimization (IACO) algorithm to perform feature selection. After this step, the selected features were used to build different SVM models. They compared traditional SVM to SVM coupled with four feature selection algorithms, including SVM-IACO. When distinguishing between bipolar disorder and MDD, SVM-IACO outperformed all other models with an accuracy of 80.19% based on 22 features. Other models had accuracies ranging from 62.37–78.21% with 25–48 features (Erguzel et al., 2015). In another study, Erguzel et al. used Cordance, a quantitative electroencepalographic method, to differentiate bipolar disorder from unipolar depression subjects. After using the Particle Swarm Optimization (PSO) algorithm for feature selection, authors create an Artificial Neural Networks (ANN) model with the selected EEG alpha and theta frequencies, representing an improvement of the model with no feature selection (accuracies of 89.89 and 73.03%, respectively) (Erguzel et al., 2016). These studies highlight the importance of improving models through appropriate feature selection, thus removing variables that may be causing noise in the model (Erguzel et al., 2015). IACO and PSO algorithms are based on insects and animal's social behavior, respectively.

In another article of the same group, authors used event-related potentials derived from EEG recordings and auditory-evoked responses to select endophenotype candidates for schizophrenia. Using SVM and these endophenotype candidates, the authors differentiated schizophrenia from HC and bipolar disorder. The optimal model classification of schizophrenia versus bipolar disorder had an accuracy of 72%, 74% sensitivity and 70% specificity (Johannesen et al., 2013). Khodayari-Rostamabad et al. used EEG data and a mixture of factor analysis (MFA) model to classify bipolar disorder, MDD, schizophrenia and HC subjects. The authors obtained 92.7% accuracy while differentiating between bipolar disorder and MDD. The analysis was repeated with another three algorithms, including SVM, but MFA accuracy exceeded all of them. Classification between bipolar disorder versus schizophrenia and bipolar disorder versus HC was not performed in this study (Khodayari-Rostamabad et al., 2010).

3.1.5. Neuropsychological tests

Similar to the studies using neuroimaging, several studies utilized machine learning algorithms and neuropsychological measures to identify patients with bipolar disorder. For instance, Wu and colleagues studied whether neurocognitive abnormalities can objectively identify individual patients with euthymic bipolar disorder from HC (Wu et al., 2016b). The authors used the LASSO algorithm to identify individual patients with bipolar disorder with an accuracy of 71% and an AUC of 0.714. In addition, each subject was assigned a probability score, estimating whether the individual belonged to the bipolar disorder or the HC group. Patients with rapid cycling were assigned increased probability scores to belong to bipolar disorder group compared to patients without rapid cycling.

Another study used machine learning classification methods to discriminate patients with late-onset bipolar disorder (LOBD) from patients with Alzheimer's disease (AD) and HC (Besga et al., 2015). Authors combined clinical observations, neuropsychological tests, and inflammatory markers to build three machine learning models using different algorithms. They analyzed the accuracy of these variables alone and combined, achieving a higher accuracy when using them altogether. The discrimination of LOBD vs. AD patients had an accuracy of 90.26%, and the discrimination of LOBD vs. HC had an accuracy of 95.76% with the combined data. It is interesting to note that blood biomarkers alone poorly discriminated LOBD from HC (46.35–60.34%) and LOBD from AD (46.38–71.01%) in all algorithms. In contrast, neuropsychological tests achieved the greatest discrimination (89.66–91.38 and 79.71–85.51% respectively). This highlights the importance of neurocognitive evaluation and the challenge of finding clinically relevant, reliable biomarkers for psychiatric disorders.

Neuropsychological tests were also applied in another study, in which the authors developed a non-invasive video-based technique to diagnose bipolar disorder. The authors used an eye pupil detection system to monitor and track the different positions of the pupil. Additionally, the system tracked the time duration of the pupils when looking in certain directions and making decisions. Using different training and testing samples from the eye-pupil data, the authors built a SVM algorithm which distinguished bipolar disorder from HC subjects with an accuracy of 96.36%. Although these results are impressive, the article focused on technical explanations and had some limitations. For instance, the authors did not explain how the bipolar disorder diagnosis was assessed or determined, and there was no sociodemographic information available to compare bipolar disorder and HC groups (Akinci et al., 2013).

3.1.6. Serum biomarkers

Haenisch and colleagues used a proteomic based approach in order to develop a potential biomarker blood panel for diagnosing bipolar disorder. In the discovery stage, they meta-analyzed eight case-control studies and used LASSO algorithm to obtain the most relevant biomarkers. After that, they validated this panel in a new sample of bipolar disorder patients and then applied the panel to bipolar disorder, MDD, HC, schizophrenia and pre-diagnostic bipolar disorder patients from three nested case-control studies using a logistic regression model. They obtained AUCs of 0.84 when distinguishing first onset MDD from undiagnosed bipolar disorder, 0.79 when comparing bipolar disorder to HC, and 0.91 when differentiating bipolar disorder from schizophrenia subjects. We highlight that 11 out of the 20 protein analytes in the panel were related to inflammatory processes. Seven of them were pro-inflammatory, and the other four were anti-inflammatory (Haenisch et al., 2016). These inflammatory proteins are consistent with recent evidence suggesting that immune activation and inflammatory processes are involved in the genesis and course of bipolar disorder (Réus et al., 2015).

In a study of our group, peripheral biomarkers (inflammatory markers, neurotrophins and oxidative stress markers) were used to differentiate HC, bipolar disorder and schizophrenia patients, with patients individually matched by age and gender. Using a SVM algorithm, the model could distinguish schizophrenia patients from HC with 72.5% accuracy ($p < 0.05$) and bipolar disorder from HC with 77.5% accuracy ($p < 0.05$). The model, however, failed to differentiate bipolar disorder from schizophrenia subjects, with an accuracy of 49% (Pinto et al., 2017). We hypothesize that the inability to separate bipolar disorder from schizophrenia in the model may be due to shared characteristics among the disorders, such as cytokines profiles and immune-inflammatory related pathways (Goldsmith et al., 2016; Goodkind et al., 2015).

3.1.7. Commentary

Regardless of the data included in the model, it is important to notice that they can never outperform our current diagnosis system. Nevertheless, they may be an important tool to predict progression to a psychiatric disorder before its installment, improving care and prophylaxis for subjects at risk. Another application would be in the forensic setting, where neuroimaging coupled with clinical and demographic data could be helpful to identify malingering, providing more reliable diagnosis.

When a model generates low accuracies, it is pertinent to ask oneself if it was a problem of the model itself, or because the diagnosis, i.e., the classes we feed the machine with do not reflect the real complexity of the illness, and there are new potential labels that could be discovered by a data-driven approach, with better integration of the multimodal data. In this fashion, machine learning could help us realize if our clinical diagnosis is consistent or not with the neurobiological

characteristics of each affected individual (Mwangi et al., 2016).

3.2. Predicting clinical outcomes

Salvini and colleagues used demographic and clinical features, including follow-up variables, to assess depression relapse in 108 bipolar disorder patients, with a total of 86 relapse cases, achieving an accuracy of 85%, and a sensitivity of 92% (Salvini et al., 2015). These are interesting findings since if one can predict which patients will most likely relapse, then early intervention could improve prognosis. Besides episode relapse, three articles used monitoring systems for long and short-term data acquisition of autonomic measures in patients with mood disorders to predict mood changes. These measures included the interbeat interval series (RR) extracted from electrocardiogram (ECG) and respiration signals (Valenza et al., 2013); ECG, respirogram and body posture data (Valenza et al., 2014); and a number of features of heart rate variability in ECG recordings (Gentili et al., 2017). Accuracies ranged from 88 to 97%, 70.8 to 96.25% and 68.57 to 99.25%, respectively. Another study used voice features collected in phone calls to classify patients' affective states, achieving AUCs of 0.78 (depressed vs. euthymic) and 0.89 (manic/mixed vs. euthymic). These proof-of-concept and experimental protocols illustrate machine learning's potential to aid in the clinical assessment of bipolar disorder patients, as models with sufficient accuracy to monitor mood states in real time may help assess disease activity and early intervention. Although promising, most of these studies had small samples, and results, therefore, need to be interpreted with caution until adequate model validation in different settings and populations.

Two studies integrated blood gene expression data with clinical information in order to predict suicidal ideation and future hospitalizations in patients with major psychiatric disorders in four different cohorts of bipolar and other psychiatric disorders participants. Niculescu and colleagues assessed male participants and reported that the SLC4A4 biomarker predicted suicidal ideation and future hospitalization for suicidality in the first year in patients with bipolar disorder with an AUC of 93% (P-value 0.45e-6) and 70% (P-value 0.08), respectively. When using the 11 top biomarkers coupled with clinical scores, an AUC of 98% (P-value 1.19e-6) was found for predicting suicidal ideation and an AUC of 94% (P-value 0.0021) for predicting future hospitalizations in the same population (Niculescu et al., 2015). Levey and colleagues repeated the same methodology with female participants, however with smaller samples, founding AUCs of 0.82 (*p*-value 0.003) and 0.78 (*p*-value 0.032) to predict suicidal ideation and future hospitalizations for suicidality while combining clinical scales and biomarkers. Results are not discriminated to the bipolar disorder subpopulation, presumably because of the smaller samples of participants. In a study of our group, we tested a set of machine learning algorithms coupled with clinical and demographic variables in an attempt to identify a clinical signature of suicidality in patients with mood disorders, including those with bipolar disorder (Passos et al., 2016a). The study presented an accuracy of 72% in predicting suicide attempts. Prior hospitalizations for depression, comorbid post-traumatic stress disorder, cocaine dependence, and history of psychotic symptoms were the most robust variables for the model.

Finally, two studies assessed treatment outcomes. In a case control-study with more than 5700 patients undergoing lithium treatment, authors achieved an AUC of 0.81 in a predictive tool for the risk development of renal insufficiency among lithium-treated patients. Nevertheless, due to the multifactorial basis of renal failure, some of the identified risks were not explained by lithium intake alone. Another important limitation was the relatively short follow-up duration (Castro et al., 2016). Regardless of these limitations, these findings indicate the possibility of stratifying risk of renal failure in this population. Moreover, it suggests that risk stratification can be expanded to other treatments and interventions. We found only one study assessing treatment response in bipolar disorder patients. Wade et al. developed a SVM model to predict electroconvulsive therapy (ECT) response in 53 patients with a major depressive episode, being 8 of them diagnosed with bipolar depression. Data used was comprised of sMRI morphometric features of caudate, putamen, pallidum and nucleus accumbens combined with three mood scales. They obtained an 89% accuracy to predict treatment response, although bipolar disorder patients were a small part of the sample, thus limiting interpretation in this particular group (Wade et al., 2016). The trajectory of bipolar disorder is largely variable, and it seems that a subset of patients will develop a more pernicious course associated with suicide attempts, relapse, and treatment refractoriness (Costa et al., 2015; Passos et al., 2016c). Although these outcomes are largely reported in the available literature, predicting them was, until recently, an elusive goal.

3.3. Data clustering using unsupervised and semi-supervised machine learning

Five studies used unsupervised learning to find natural groupings or clusters of patients that may correspond to biologically relevant phenotypes. Wahlund et al. used principle component analysis to cluster patients with unipolar depression based on clinical and biological characteristics. Five of the 28 patients initially included presented manic or hypomanic symptoms in the 15 years follow-up. Most of these manic patients belonged to a cluster with significantly higher levels of MAO activity and melatonin, identified prior to the follow-up. The bipolar disorder patients differed from one of the unipolar depression clusters in psychomotor symptoms, and from another unipolar depression cluster in post-dexamethasone serum cortisol level. This study demonstrates the potential to identify patients with unipolar depression who will convert to bipolar disorder by the use of biomarkers. Unfortunately, these serum markers are highly variable and labor intensive to acquire. Additionally, the follow-up time of 15 years may have been insufficient to identify all patients who eventually converted to bipolar disorder (Wahlund et al., 1998).

Three recent studies assessed patients with bipolar disorder using unsupervised machine learning to address different questions. Hall and colleagues used K-means algorithm to identify neurophysiologic profiles in subgroups of patients, relatives and controls in two independent samples. In their first analysis, authors included all subjects obtaining three groups: globally impaired, sensory processing, and high cognitive, with most of the patients being classified in the globally impaired group ($p < 0.001$). In the second analysis, however, schizophrenia and bipolar disorder patients did not have distinct neurophysiological profiles. The authors only included bipolar disorder patients with psychotic symptoms, a profile that may be more related to patients with schizophrenia than non-psychotic bipolar disorder, which can explain this result (Hall et al., 2012).

Wu and colleagues identified two bipolar disorder phenotypes distinct from DSM-IV subtypes in a semi-supervised model using brain diffusion weighted imaging and neurocognitive data from the computerized Cambridge Neurocognitive Test Automated Battery (CANTAB) (Wu et al., 2016a). These data driven phenotypes are in consonance with the cognitive profiles of the RdOC constructs, suggesting that there is enormous potential for these descriptive models to improve our diagnostic paradigm into a dimensional approach rather than a categorical one. In another study, Bansal and colleagues created an algorithm that identified different types of neuropsychiatric illnesses such as bipolar disorder, attention-deficit/hyperactivity disorder, schizophrenia, and Tourette syndrome (Bansal et al., 2012), also using weighted imaging in a semi-supervised model. This model identified neuropsychiatric illnesses based on brain morphologic characteristics with high sensitivity and specificity (100% and 96.4% respectively in bipolar disorder vs. HC, and 99.99% and 100% respectively in bipolar disorder vs. schizophrenia).

4. Conclusion

The present study showed that neuroimaging studies, besides helping us to understand better the pathophysiology of psychiatric disorders, may also help differentiate bipolar disorder from healthy controls and other psychiatric diagnosis. That holds the potential to help in addressing the problem of misdiagnosis and diagnostic delay in BD. In the present study, we also performed a meta-analysis of diagnostic accuracy neuroimaging studies that included articles comparing subjects with bipolar disorder with healthy controls.

Machine learning techniques may also be used to assess individuals at risk, such as bipolar offspring, transforming data into applicable information about the individual risk of having a future diagnosis. In this sense, machine learning may allow to develop personalized interventions to prevent the transition from prodromes to full-blown illness among high-risk patients. Once diagnosis is established, machine learning may help to predict to which treatment patients are more likely to benefit from. Prediction of treatment response may potentially reduce the duration of mood episodes and help to tailor the maintenance treatment.

In addition, the present study showed that outcomes, such as suicide, hospitalizations and episode relapse could also be predicted with reasonable accuracy by using machine learning models. Emerging data also suggests that integration with mobile devices and social media could prove a useful resource to recognize a subject mood state, allowing the clinician to be advised prior to the onset of a mood episode in real time.

However, most of the included studies are still in a proof-of-concept phase, with small sample sizes and lack of adequate external validation. For instance, structural neuroimaging studies are often performed in subjects already diagnosed with a psychiatric disorder with years of progress and medication use, two confounding factors. It is hard to determine whether structural brain alterations are the result, risk factors or the cause of disease. Despite that, neuroimaging coupled with machine learning may serve for multiple purposes. Not only it can be used as a tool for predicting individuals at risk, but it can also be applied to patients with chronic disease to study cognitive impairments, neuroprogression status and to establish disease phenotypes.

Another limitation is the lack of population studies to validate most of the described machine learning algorithms. Future studies on big data analytics and machine learning should focus on the following: validation of the predictive algorithms using new and independent datasets; proposal and implementation of new models to predict important clinical outcomes such as mood episode relapse and cognitive impairment; and development of web-based calculators and medical devices to allow translation of these predictive models into the clinical practice. We postulate that highly accurate predictive models will support important clinical decisions such as selection of treatment options, preventive strategies, and prognosis orientations.

Finally, preliminary findings and proof-of-concept studies, along with future studies, may help us shift from diagnostic categories based on patients' subjective description of symptoms and clinical observations, to an intel-based integrative paradigm. Machine learning techniques may help to define clusters of patients who share similar characteristics in a more complex level than our current classification systems allows for. This could contribute to generate better staging systems, allowing us to detect subgroups of patients with similar outcomes. Another great challenge for the future is to make these complex features and algorithms available to clinicians in a practical and applicable manner, turning big data analytics and machine learning models into real benefits for patients.

Conflict of interest statements

Diego Librenza-Garcia, Bruno Jaskulski Kotzian, Jessica Yang, Benson Mwangi, Bo Cao, Luiza Nunes Pereira Lima, Mariane Bagatin Bermudez, Manuela Vianna Boeira and Ives Cavalcante Passos reported 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 Astra Zeneca, Eli Lilly, Janssen and Servier; and has served as a consultant for Servier.

Author's contributions

Diego Librenza-Garcia, Luiza Nunes Pereira Lima and Ives Cavalcante Passos participated in the literature search, writing, and in the approval of the final manuscript. Bruno Jaskulski Kotzian, Jessica Yang, Benson Mwangi, Bo Cao, Mariane Bagatin Bermudez, Manuela Vianna Boeira and Flávio Kapczinski participated in the writing and in the approval of the final manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.neubiorev.2017.07.004.

References

Acikel, C., Aydin Son, Y., Celik, C., Gul, H., 2016. Evaluation of novel candidate variations and their interactions related to bipolar disorders: analysis of GWAS data. Neuropsychiatr. Dis. Treat. 12, 2997–3004. http://dx.doi.org/10.2147/NDT.S112558.

Akinci, G., Polat, E., Koçak, O.M., 2013. A video-based eye pupil detection system for diagnosing bipolar disorder. Turk. J. Electr. Eng. Comput. Sci. 21, 2367–2377. http://dx.doi.org/10.3906/elk-1204-63.

Almeida, J.R.C., Mourao-Miranda, J., Aizenstein, H.J., Versace, A., Kozel, F.A., Lu, H., Marquand, A., LaBarbara, E.J., Brammer, M., Trivedi, M., Kupfer, D.J., Phillips, M.L., 2013. Pattern recognition analysis of anterior cingulate cortex blood flow to classify depression polarity. Br. J. Psychiatry 203, 310–311. http://dx.doi.org/10.1192/bjp.bp.112.122838.

Anticevic, A., Cole, M.W., Repovs, G., Murray, J.D., Brumbaugh, M.S., Winkler, A.M., Savic, A., Krystal, J.H., Pearlson, G.D., Glahn, D.C., 2014. Characterizing thalamocortical disturbances in schizophrenia and bipolar illness. Cereb. Cortex 24, 3116–3130. http://dx.doi.org/10.1093/cercor/bht165.

Arribas, J.I., Calhoun, V.D., Adali, T., 2010. Automatic bayesian classification of healthy controls, bipolar disorder, and schizophrenia using intrinsic connectivity maps from fMRI data. IEEE Trans. Biomed. Eng. 57, 2850–2860. http://dx.doi.org/10.1109/TBME.2010.2080679.

Bansal, R., Staib, L.H., Laine, A.F., Hao, X., Xu, D., Liu, J., Weissman, M., Peterson, B.S., 2012. Anatomical brain images alone can accurately diagnose chronic neuropsychiatric illnesses. PLoS One 7, e50698. http://dx.doi.org/10.1371/journal.pone.0050698.

Berk, M., Kapczinski, F., Andreazza, A.C., Dean, O.M., Giorlando, F., Maes, M., Yücel, M., Gama, C.S., Dodd, S., Dean, B., Magalhães, P.V.S., Amminger, P., McGorry, P., Malhi, G.S., 2011. Pathways underlying neuroprogression in bipolar disorder: focus on inflammation, oxidative stress and neurotrophic factors. Neurosci. Biobehav. Rev. 35, 804–817. http://dx.doi.org/10.1016/j.neubiorev.2010.10.001.

Besga, A., Termenon, M., Graña, M., Echeveste, J., Pérez, J.M., Gonzalez-Pinto, A., 2012. Discovering Alzheimer's disease and bipolar disorder white matter effects building computer aided diagnostic systems on brain diffusion tensor imaging features. Neurosci. Lett. 520, 71–76. http://dx.doi.org/10.1016/j.neulet.2012.05.033.

Besga, A., Gonzalez, I., Echeburua, E., Savio, A., Ayerdi, B., Chyzhyk, D., Madrigal, J.L.M., Leza, J.C., Graña, M., Gonzalez-Pinto, A.M., 2015. Discrimination between Alzheimer's disease and late onset bipolar disorder using multivariate analysis. Front. Aging Neurosci. 7, 1–9. http://dx.doi.org/10.3389/fnagi.2015.00231.

Castro, V.M., Roberson, A.M., McCoy, T.H., Wiste, A., Cagan, A., Smoller, J.W., Rosenbaum, J.F., Ostacher, M., Perlis, R.H., 2016. Stratifying risk for renal insufficiency among lithium-treated patients: an electronic health record study. Neuropsychopharmacology 41, 1138–1143. http://dx.doi.org/10.1038/npp.2015.254.

Chen, Y., Storrs, J., Tan, L., Mazlack, L.J., Lee, J., Lu, L.J., 2014. Detecting brain structural changes as biomarker from magnetic resonance images using a local feature based SVM approach. J. Neurosci. Methods 221, 22–31. http://dx.doi.org/10.1016/j.jneumeth.2013.09.001.

Chuang, L.-C., Kuo, P.-H., 2017. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. Sci. Rep. 7, 39943. http://dx.doi.org/10.1038/srep39943.

Costa, L.D.S., Alencar Á, P., Neto, P.J.N., dos Santos, M.do S.V., da Silva, C.G.L., Pinheiro, S.D.F.L., Teixeira Silveira, R., Bianco, B.A.V., Pinheiro Júnior, R.F.F., de Lima, M.A.P., Reis, A.O.A., Neto, M.L.R., 2015. Risk factors for suicide in bipolar disorder: a systematic review. J. Affect. Disord. 170, 237–254. http://dx.doi.org/10.1016/j.jad.2014.09.003.

Costafreda, S.G., Fu, C.H., Picchioni, M., Toulopoulou, T., McDonald, C., Kravariti, E., Walshe, M., Prata, D., Murray, R.M., McGuire, P.K., 2011. Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. BMC Psychiatry 11, 18. http://dx.doi.org/10.1186/1471-244X-11-18.

Crump, C., Sundquist, K., Winkleby, M.A., Sundquist, J., 2013. Comorbidities and mortality in bipolar disorder. JAMA Psychiatry 70, 931. http://dx.doi.org/10.1001/jamapsychiatry.2013.1394.

Dmitrzak-Weglarz, M.P., Pawlak, J.M., Maciukiewicz, M., Moczko, J., Wilkosc, M., Leszczynska-Rodziewicz, A., Zaremba, D., Hauser, J., 2015. Clock gene variants differentiate mood disorders. Mol. Biol. Rep. 42, 277–288. http://dx.doi.org/10.1007/s11033-014-3770-9.

Du, Y., Pearlson, G.D., Liu, J., Sui, J., Yu, Q., He, H., Castro, E., Calhoun, V.D., 2015. A group ICA based framework for evaluating resting fMRI markers when disease categories are unclear: application to schizophrenia, bipolar, and schizoaffective disorders. Neuroimage 122, 272–280. http://dx.doi.org/10.1016/j.neuroimage.2015.07.054.

Erguzel, T.T., Tas, C., Cebi, M., 2015. A wrapper-based approach for feature selection and classification of major depressive disorder–bipolar disorders. Comput. Biol. Med. 64, 127–137. http://dx.doi.org/10.1016/j.compbiomed.2015.06.021.

Erguzel, T.T., Sayar, G.H., Tarhan, N., 2016. Artificial intelligence approach to classify unipolar and bipolar depressive disorders. Neural Comput. Appl. 27, 1607–1616. http://dx.doi.org/10.1007/s00521-015-1959-z.

Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E.M., Winther, O., Bardram, J.E., Kessing, L.V., 2016. Voice analysis as an objective state marker in bipolar disorder. Transl. Psychiatry 6, e856. http://dx.doi.org/10.1038/tp.2016.123.

Frangou, S., Dima, D., Jogia, J., 2017. Towards person-centered neuroimaging markers for resilience and vulnerability in Bipolar Disorder. Neuroimage 145, 230–237. http://dx.doi.org/10.1016/j.neuroimage.2016.08.066.

Fung, G., Deng, Y., Zhao, Q., Li, Z., Qu, M., Li, K., Zeng, Y.-W., Jin, Z., Ma, Y.-T., Yu, X., Wang, Z.-R., Shum, D.H.K., Chan, R.C.K., 2015. Distinguishing bipolar and major depressive disorders by brain structural morphometry: a pilot study. BMC Psychiatry 15, 298. http://dx.doi.org/10.1186/s12888-015-0685-5.

Gentili, C., Valenza, G., Nardelli, M., Lanatà, A., Bertschy, G., Weiner, L., Mauri, M., Scilingo, E.P., Pietrini, P., 2017. Longitudinal monitoring of heartbeat dynamics predicts mood changes in bipolar patients: a pilot study. J. Affect. Disord. 209, 30–38. http://dx.doi.org/10.1016/j.jad.2016.11.008.

Gitlin, M.J., Swendsen, J., Heller, T.L., Hammen, C., 1995. Relapse and impairment in bipolar disorder. Am. J. Psychiatry 152, 1635–1640. http://dx.doi.org/10.1176/ajp.152.11.1635.

Goldberg, J.F., Harrow, M., Whiteside, J.E., 2001. Risk for bipolar illness in patients initially hospitalized for unipolar depression. Am. J. Psychiatry 158, 1265–1270. http://dx.doi.org/10.1176/appi.ajp.158.8.1265.

Goldsmith, D.R., Rapaport, M.H., Miller, B.J., 2016. A meta-analysis of blood cytokine network alterations in psychiatric patients: comparisons between schizophrenia, bipolar disorder and depression. Mol. Psychiatry 1–14. http://dx.doi.org/10.1038/mp.2016.3.

Goodkind, M., Eickhoff, S.B., Oathes, D.J., Jiang, Y., Chang, A., Jones-Hagata, L.B., Ortega, B.N., Zaiko, Y.V., Roach, E.L., Korgaonkar, M.S., Grieve, S.M., Galatzer-Levy, I., Fox, P.T., Etkin, A., 2015. Identification of a common neurobiological substrate for mental illness. JAMA Psychiatry 72, 305–315. http://dx.doi.org/10.1001/jamapsychiatry.2014.2206.

Greenhalgh, T., Howick, J., Maskrey, N., 2014. Evidence based medicine: a movement in crisis? BMJ 348, g3725. http://dx.doi.org/10.1136/bmj.g3725.

Grotegerd, D., Suslow, T., Bauer, J., Ohrmann, P., Arolt, V., Stuhrmann, A., Heindel, W., Kugel, H., Dannlowski, U., 2013. Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: a pilot study. Eur. Arch. Psychiatry Clin. Neurosci. 263, 119–131. http://dx.doi.org/10.1007/s00406-012-0329-4.

Grotegerd, D., Stuhrmann, A., Kugel, H., Schmidt, S., Redlich, R., Zwanzger, P., Rauch, A.V., Heindel, W., Zwitserlood, P., Arolt, V., Suslow, T., Dannlowski, U., 2014. Amygdala excitability to subliminally presented emotional faces distinguishes unipolar and bipolar depression: an fMRI and pattern classification study. Hum. Brain Mapp. 35, 2995–3007. http://dx.doi.org/10.1002/hbm.22380.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Haenisch, F., Cooper, J.D., Reif, A., Kittel-Schneider, S., Steiner, J., Leweke, F.M., Rothermundt, M., van Beveren, N.J.M., Crespo-Facorro, B., Niebuhr, D.W., Cowan, D.N., Weber, N.S., Yolken, R.H., Penninx, B.W.J.H., Bahn, S., 2016. Towards a blood-based diagnostic panel for bipolar disorder. Brain. Behav. Immun. 52, 49–57. http://dx.doi.org/10.1016/j.bbi.2015.10.001.

Hajek, T., Cooke, C., Kopecek, M., Novak, T., Hoschl, C., Alda, M., 2015. Using structural MRI to identify individuals at genetic risk for bipolar disorders: a 2-cohort, machine learning study. J. Psychiatry Neurosci. 40, 316–324. http://dx.doi.org/10.1503/jpn.140142.

Hall, M.-H., Smoller, J.W., Cook, N.R., Schulze, K., Hyoun Lee, P., Taylor, G., Bramon, E., Coleman, M.J., Murray, R.M., Salisbury, D.F., Levy, D.L., 2012. Patterns of deficits in brain function in bipolar disorder and schizophrenia: a cluster analytic study. Psychiatry Res. 200, 272–280. http://dx.doi.org/10.1016/j.psychres.2012.07.052.

Huys, Q.J.M., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19, 404–413. http://dx.doi.org/10.1038/nn.4238.

Jie, N., Zhu, M., Ma, X., Osuch, E.A., Wammes, M., Theberge, J., Li, H., Zhang, Y., Jiang, T.-Z., Sui, J., Calhoun, V.D., 2015. Discriminating bipolar disorder from major depression based on SVM-FoBa: efficient feature selection with multimodal brain imaging data. IEEE Trans. Auton. Ment. Dev. 7, 320–331. http://dx.doi.org/10.1109/TAMD.2015.2440298.

Johannesen, J.K., O'Donnell, B.F., Shekhar, A., McGrew, J.H., Hetrick, W.P., 2013. Diagnostic specificity of neurophysiological endophenotypes in schizophrenia and bipolar disorder. Schizophr. Bull. 39, 1219–1229. http://dx.doi.org/10.1093/schbul/sbs093.

Kapczinski, N.S., Mwangi, B., Cassidy, R.M., Librenza-Garcia, D., Bermudez, M.B., Kauer-Sant'anna, M., Kapczinski, F., Passos, I.C., 2016. Neuroprogression and illness trajectories in bipolar disorder. Expert Rev. Neurother. 0, 1–9. http://dx.doi.org/10.1080/14737175.2017.1240615.

Kaufmann, T., Alnæs, D., Brandt, C.L., Doan, N.T., Kauppi, K., Bettella, F., Lagerberg, T.V., Berg, A.O., Djurovic, S., Agartz, I., Melle, I.S., Ueland, T., Andreassen, O.A., Westlye, L.T., 2017. Task modulations and clinical manifestations in the brain functional connectome in 1615 fMRI datasets. Neuroimage 147, 243–252. http://dx.doi.org/10.1016/j.neuroimage.2016.11.073.

Khodayari-Rostamabad, A., Reilly, J.P., Hasey, G., de Bruin, H., MacCrimmon, D., 2010. Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE. pp. 4006–4009. http://dx.doi.org/10.1109/IEMBS.2010.5627998.

Koutsouleris, N., Meisenzahl, E.M., Borgwardt, S., Riecher-Rössler, A., Frodl, T., Kambeitz, J., Köhler, Y., Falkai, P., Möller, H.-J., Reiser, M., Davatzikos, C., 2015. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. Brain 138, 2059–2073. http://dx.doi.org/10.1093/brain/awv111.

Lantz, B., 2015. Machine Learning with R, Second Edition. Packt Publishing. Cambridge University Press, Cambridge.

Levey, D.F., Niculescu, E.M., Le-Niculescu, H., Dainton, H.L., Phalen, P.L., Ladd, T.B., Weber, H., Belanger, E., Graham, D.L., Khan, F.N., Vanipenta, N.P., Stage, E.C., Ballew, A., Yard, M., Gelbart, T., Shekhar, A., Schork, N.J., Kurian, S.M., Sandusky, G.E., Salomon, D.R., Niculescu, A.B., 2016. Towards understanding and predicting suicidality in women: biomarkers and clinical risk assessment. Mol. Psychiatry 21, 768–785. http://dx.doi.org/10.1038/mp.2016.31.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gotzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 339, b2700. http://dx.doi.org/10.1136/bmj.b2700.

Lish, J.D., Dime-Meenan, S., Whybrow, P.C., Price, R.A., Hirschfeld, R.M.A., 1994. The national depressive and manic-depressive association (DMDA) survey of bipolar members. J. Affect. Disord. 31, 281–294. http://dx.doi.org/10.1016/0165-0327(94)90104-X.

Mathers, C.D., Iburg, K.M., Begg, S., 2006. Adjusting for dependent comorbidity in the calculation of healthy life expectancy. Popul. Health Metr. 4, 4. http://dx.doi.org/10.1186/1478-7954-4-4.

Merikangas, K.R., Akiskal, H.S., Angst, J., Greenberg, P.E., Hirschfeld, R.M.A., Petukhova, M., Kessler, R.C., 2007. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. Arch. Gen. Psychiatry 64, 543. http://dx.doi.org/10.1001/archpsyc.64.5.543.

Mourão-Miranda, J., Almeida, J.R., Hassel, S., de Oliveira, L., Versace, A., Marquand, A.F., Sato, J.R., Brammer, M., Phillips, M.L., 2012. Pattern recognition analyses of brain activation elicited by happy and neutral faces in unipolar and bipolar depression. Bipolar Disord. 14, 451–460. http://dx.doi.org/10.1111/j.1399-5618.2012.01019.x.

Mwangi, B., Ebmeier, K.P., Matthews, K., Douglas Steele, J., 2012. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain 135, 1508–1521. http://dx.doi.org/10.1093/brain/aws084.

Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244. http://dx.doi.org/10.1007/s12021-013-9204-3.

Mwangi, B., Wu, M., Cao, B., Passos, I.C., Lavagnino, L., Keser, Z., Zunta-Soares, G.B., Hasan, K.M., Kapczinski, F., Soares, J.C., 2016. Individualized prediction and clinical staging of bipolar disorders using neuroanatomical biomarkers. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1, 186–194. http://dx.doi.org/10.1016/j.bpsc.2016.01.001.

Niculescu, A.B., Levey, D.F., Phalen, P.L., Le-Niculescu, H., Dainton, H.D., Jain, N., Belanger, E., James, A., George, S., Weber, H., Graham, D.L., Schweitzer, R., Ladd, T.B., Learman, R., Niculescu, E.M., Vanipenta, N.P., Khan, F.N., Mullen, J., Shankar, G., Cook, S., Humbert, C., Ballew, A., Yard, M., Gelbart, T., Shekhar, A., Schork, N.J., Kurian, S.M., Sandusky, G.E., Salomon, D.R., 2015. Understanding and predicting suicidality using a combined genomic and clinical risk assessment approach. Mol. Psychiatry 20, 1266–1285. http://dx.doi.org/10.1038/mp.2015.112.

Nordentoft, M., 2011. Absolute risk of suicide after first hospital contact in mental disorder. Arch. Gen. Psychiatry 68, 1058. http://dx.doi.org/10.1001/archgenpsychiatry.2011.113.

Passos, I.C., Mwangi, B., Kapczinski, F., 2016a. Big data analytics and machine learning: 2015 and beyond. Lancet Psychiatry 3, 13–15. http://dx.doi.org/10.1016/S2215-0366(15)00549-0.

Passos, I.C., Mwangi, B., Vieta, E., Berk, M., Kapczinski, F., 2016b. Areas of controversy in neuroprogression in bipolar disorder. Acta Psychiatr. Scand. 134, 91–103. http://dx.doi.org/10.1111/acps.12581.

Passos, I.C., Mwangi, B., Cao, B., Hamilton, J.E., Wu, M.-J., Zhang, X.Y., Zunta-Soares, G.B., Quevedo, J., Kauer-Sant'Anna, M., Kapczinski, F., Soares, J.C., 2016c. Identifying a clinical signature of suicidality among patients with mood disorders: a pilot study using a machine learning approach. J. Affect. Disord. 193, 109–116. http://dx.doi.org/10.1016/j.jad.2015.12.066.

Petcharunpaisan, S., Ramalho, J., Castillo, M., 2010. Arterial spin labeling in neuroimaging. World J. Radiol. 2, 384–398. http://dx.doi.org/10.4329/wjr.v2.i10.384.

Pinto, J.V., Passos, I.C., Gomes, F., Reckziegel, R., Kapczinski, F., Mwangi, B., Kauer-Sant'Anna, M., 2017. Peripheral biomarker signatures of bipolar disorder and schizophrenia: a machine learning approach. Schizophr. Res. 6, 2016–2018. http://dx.doi.org/10.1016/j.schres.2017.01.018.

Pirooznia, M., Seifuddin, F., Judy, J., Mahon, P.B., Potash, J.B., Zandi, P.P., 2012. Data mining approaches for genome-wide association of mood disorders. Psychiatr. Genet. 22, 55–61. http://dx.doi.org/10.1097/YPG.0b013e32834dc40d.

Réus, G.Z., Fries, G.R., Stertz, L., Badawy, M., Passos, I.C., Barichello, T., Kapczinski, F., Quevedo, J., 2015. The role of inflammation and microglial activation in the pathophysiology of psychiatric disorders. Neuroscience 300, 141–154. http://dx.doi.org/10.1016/j.neuroscience.2015.05.018.

Redlich, R., Almeida, J.J.R., Grotegerd, D., Opel, N., Kugel, H., Heindel, W., Arolt, V., Phillips, M.L., Dannlowski, U., 2014. Brain morphometric biomarkers distinguishing unipolar and bipolar depression. JAMA Psychiatry 71, 1222. http://dx.doi.org/10.1001/jamapsychiatry.2014.1100.

Reitsma, J.B., Glas, A.S., Rutjes, A.W.S., Scholten, R.J.P.M., Bossuyt, P.M., Zwinderman, A.H., 2005. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J. Clin. Epidemiol. 58, 982–990. http://dx.doi.org/10.1016/j.jclinepi.2005.02.022.

Rive, M.M., Redlich, R., Schmaal, L., Marquand, A.F., Dannlowski, U., Grotegerd, D., Veltman, D.J., Schene, A.H., Ruhé, H.G., 2016. Distinguishing medication-free subjects with unipolar disorder from subjects with bipolar disorder: state matters. Bipolar Disord. 18, 612–623. http://dx.doi.org/10.1111/bdi.12446.

Roberts, G., Lord, A., Frankland, A., Wright, A., Lau, P., Levy, F., Lenroot, R.K., Mitchell, P.B., Breakspear, M., 2016. Functional dysconnection of the inferior frontal gyrus in young people with bipolar disorder or at genetic high risk. Biol. Psychiatry 1–10. http://dx.doi.org/10.1016/j.biopsych.2016.08.018.

Rocha-Rego, V., Jogia, J., Marquand, A.F., Mourao-Miranda, J., Simmons, A., Frangou, S., 2014. Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach. Psychol. Med. 44, 519–532. http://dx.doi.org/10.1017/S0033291713001013.

Sacchet, M.D., Livermore, E.E., Iglesias, J.E., Glover, G.H., Gotlib, I.H., 2015. Subcortical volumes differentiate major depressive disorder, bipolar disorder, and remitted major depressive disorder. J. Psychiatr. Res. 68, 91–98. http://dx.doi.org/10.1016/j.jpsychires.2015.06.002.

Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S., 1996. Evidence based medicine: what it is and what it isn't. BMJ 312, 71–72. http://dx.doi.org/10.1136/bmj.312.7023.71.

Salvini, R., da Silva Dias, R., Lafer, B., Dutra, I., 2015. A multi-relational model for depression relapse in patients with bipolar disorder. Stud. Heal. Technol. Inf. 216, 741–745. http://dx.doi.org/10.3233/978-1-61499-564-7-741.

Schnack, H.G., Nieuwenhuis, M., van Haren, N.E.M., Abramovic, L., Scheewe, T.W., Brouwer, R.M., Hulshoff Pol, H.E., Kahn, R.S., 2014. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. Neuroimage 84, 299–306. http://dx.doi.org/10.1016/j.neuroimage.2013.08.053.

Serpa, M.H., Ou, Y., Schaufelberger, M.S., Doshi, J., Ferreira, L.K., Machado-Vieira, R., Menezes, P.R., Scazufca, M., Davatzikos, C., Busatto, G.F., Zanetti, M.V., 2014. Neuroanatomical classification in a population-Based sample of psychotic major depression and bipolar I disorder with 1 year of diagnostic stability. Biomed. Res. Int. 2014, 1–9. http://dx.doi.org/10.1155/2014/706157.

Struyf, J., Dobrin, S., Page, D., 2008. Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. BMC Genom. 9, 531. http://dx.doi.org/10.1186/1471-2164-9-531.

Sun, Y., Todorovic, S., Goodison, S., 2010. Local-learning-based feature selection for high-dimensional data analysis. Pami 32, 1610–1626. http://dx.doi.org/10.1109/TPAMI.2009.190.

Tohka, J., Moradi, E., Huttunen, H., 2016. Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics 14, 279–296. http://dx.doi.org/10.1007/s12021-015-9292-3.

Valenza, G., Gentili, C., Lanatà, A., Scilingo, E.P., 2013. Mood recognition in bipolar patients through the PSYCHE platform: preliminary evaluations and perspectives. Artif. Intell. Med. 57, 49–58. http://dx.doi.org/10.1016/j.artmed.2012.12.001.

Valenza, G., Nardelli, M., Lanata, A., Gentili, C., Bertschy, G., Paradiso, R., Scilingo, E.P., 2014. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. IEEE J. Biomed. Heal. Inform. 18, 1625–1635. http://dx.doi.org/10.1109/JBHI.2013.2290382.

Wade, B.S.C., Joshi, S.H., Njau, S., Leaver, A.M., Vasavada, M., Woods, R.P., Gutman, B.A., Thompson, P.M., Espinoza, R., Narr, K.L., 2016. Effect of electroconvulsive therapy on striatal morphometry in major depressive disorder. Neuropsychopharmacology 41, 2481–2491. http://dx.doi.org/10.1038/npp.2016.48.

Wahlund, B., Grahn, H., Sääf, J., Wetterberg, L., 1998. Affective disorder subtyped by psychomotor symptoms, monoamine oxidase, melatonin and cortisol: identification of patients with latent bipolar disorder. Eur. Arch. Psychiatry Clin. Neurosci. 248, 215–224. http://dx.doi.org/10.1007/s004060050041.

Wu, M., Mwangi, B., Bauer, I.E., Passos, I.C., Sanches, M., Zunta-Soares, G.B., Meyer, T.D., Hasan, K.M., Soares, J.C., 2016a. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. Neuroimage. http://dx.doi.org/10.1016/j.neuroimage.2016.02.016.

Wu, M., Passos, I.C., Bauer, I.E., Lavagnino, L., Cao, B., Zunta-Soares, G.B., Kapczinski, F., Mwangi, B., Soares, J.C., 2016b. Individualized identification of euthymic bipolar disorder using the Cambridge Neuropsychological Test Automated Battery (CANTAB) and machine learning. J. Affect. Disord. 192, 219–225. http://dx.doi.org/10.1016/j.jad.2015.12.053.

Supplemental material

Brief overview of machine learning techniques and model validation

Machine learning algorithms can be used to build models to predict treatment response or side effects, and to perform pattern classification tasks. In supervised machine learning, the model is 'trained' using a labeled dataset (e.g. healthy vs disease). Subsequently, the 'learnt' model is applied to a 'new' or 'novel' dataset – also known as testing or validation datasets (Gollapudi, 2016; Lantz, 2015). On the other hand, unsupervised machine learning algorithms are used to discover biological groups or clusters within the sample without necessarily requiring user defined disease labels. The identified clusters or groups may correspond to new potentially homogeneous disease phenotypes or subtypes (Mwangi, Soares, & Hasan, 2014). The identified phenotypes must then be interpreted to give meaning to the resulting classes. These algorithms are unsupervised as they are 'trained' without target labels. For an instance, a supervised model may be used to distinguish between bipolar disorder and HC based on existing labeled data, while an unsupervised algorithm can learn to recognize and differentiate subtypes of bipolar disorder without any pre-defined labels. Both supervised and unsupervised learning algorithms are depicted in figure S1. Finally, it is also possible to use an unsupervised machine learning algorithm followed by a supervised algorithm – also known as a semi-supervised approach. Hence, the unsupervised algorithm may aid in dimensionality reduction before the supervised algorithm analyze the data (Hall et al., 2012; Huys, Maia, & Frank, 2016; Mwangi et al., 2014).

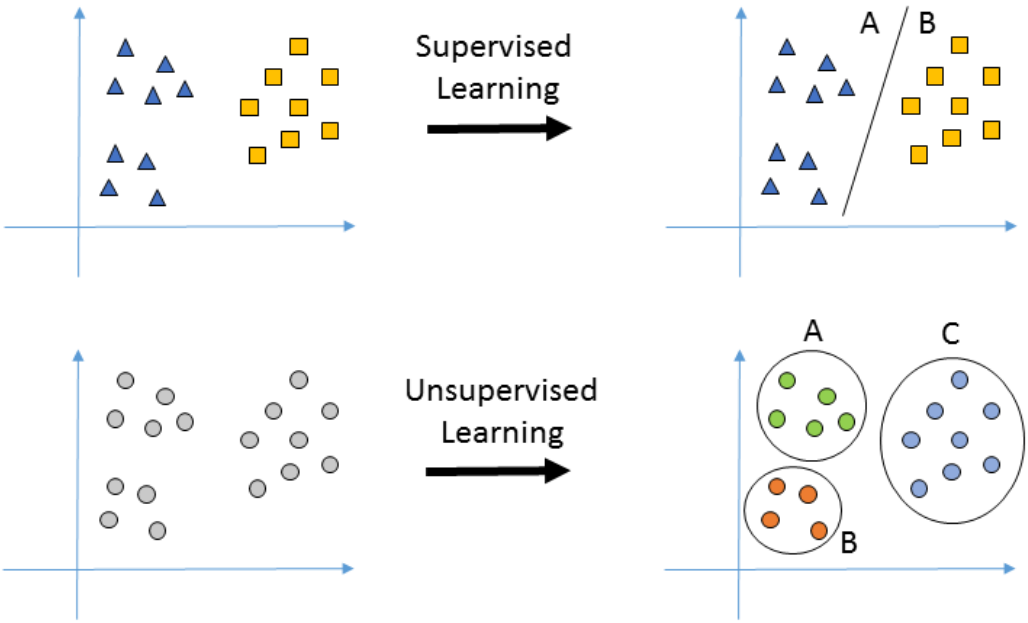


Figure S1. Supervised and unsupervised learning

In a supervised machine learning algorithm, outcome is already labeled in the training test, and the model tries to predict the outcome or class in the testing/validation set. In the example, model was already established in another data set and now it is able to differentiate the two labeled classes. In an unsupervised algorithm, the data has no outcome specified, so the algorithm will study the data set to find different groups with homogeneous characteristic (e.g. clustering.). In the figure, the algorithm found three clusters from the unlabeled data.

However, in both supervised and unsupervised approach a validation method is required. The ideal validation in the scenario of prognostic or predictive analysis using machine learning would be to perform a longitudinal study in which the subjects are followed up and the model previously developed in another sample is applied, allowing us to see how accurately the model can make its

predictions. In regard of classification studies, testing the classifier in an independent population would be preferable (Cao et al., 2015, 2016). As these approaches are not always feasible, most of the studies use cross-validation techniques. One way of doing so is to divide the data set into training and validation sets, to be used in algorithm training and validation process (Gollapudi, 2016; Lantz, 2015). Typical methods of dividing training and testing sets include K-fold (e.g. 10-fold), leave-one-out or hold-out/split-half cross-validation. For example, in a 10-fold cross-validation approach, the data is divided into 10 folds and is trained with 9-folds followed by a validation step using 1-fold. This process is repeated until all data folds are used at least once (Lantz, 2015; Struyf, Dobrin, & Page, 2008). Finally, leave-one-out cross-validation can be performed when the data set has few examples and is too small to be divided using the hold-out or k-fold approach. The leave-one-out cross-validation method leaves the data of one subject out and uses the remaining ones to build the model. Then the model is validated using the subject that was left out. This process is repeated until every subject is used to validate the model (Passos et al., 2016). These methods are illustrated in Figure S2.

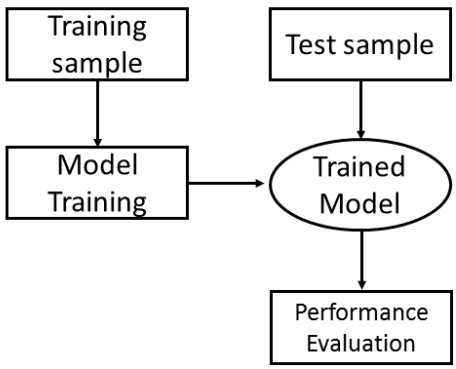


Figure S2. Validation in a new sample

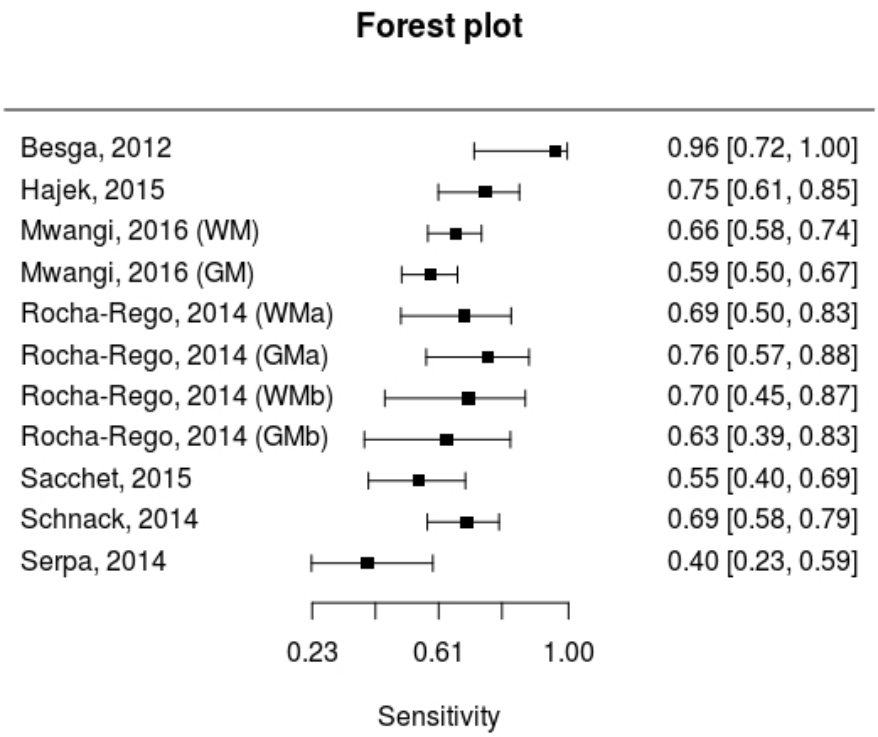
In the ideal scenario, a model is developed on one sample and then tested in another one. The performance of the model is the result of the algorithm performance in the test sample.

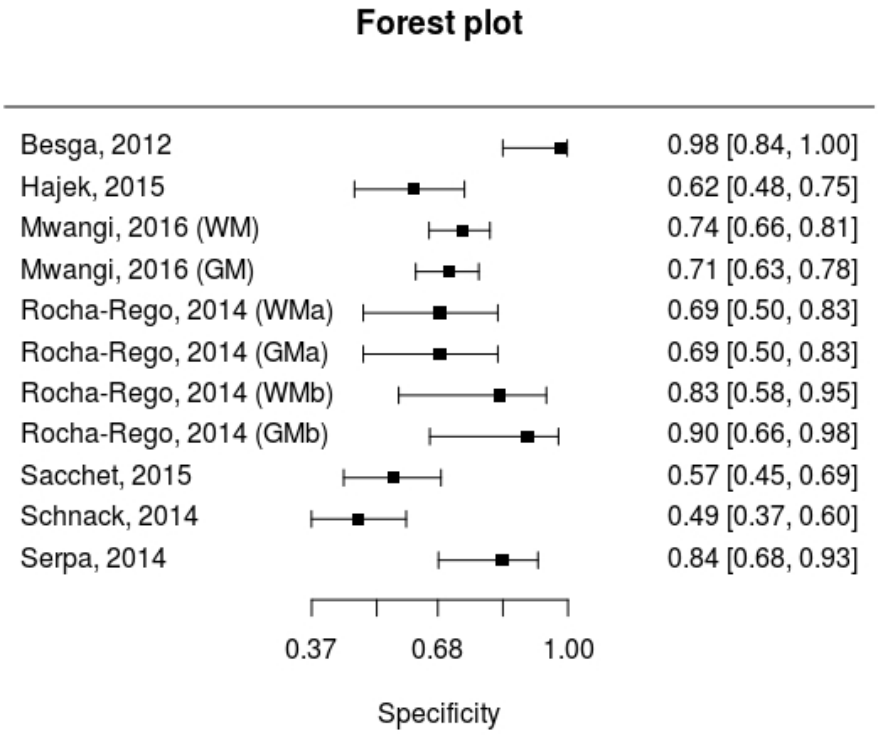
There are a wide variety of machine learning algorithms available, and their uses may depend on the purpose of the analysis, as well as intrinsic characteristics of the data. In general, supervised machine learning algorithms can be divided into two broad groups, namely Kernel-learning and penalized regression algorithms. Briefly, kernel learning algorithms use a kernel learning function (e.g. linear, polynomial) to project the data into a high-dimensional space where subjects' classes' are easily separable – an operation also known as the 'kernel trick' (Bishop, 2006). By far, Support Vector Machine (SVM) - which is a kernel learning algorithm - was the most commonly used method in the included articles in this review, followed by Gaussian Process Classifiers (GPC). An SVM algorithm establishes an optimal hyperplane that separates different classes (e.g. health vs. disease) through a sub-sample of the data also known as support vectors. Support vectors are established from the closest points to the hyperplane of the data (Cortes & Vapnik, 1995; Fung et al., 2015; Lantz, 2015). Gaussian process classifier (GPC) is based on Bayesian probability theory and uses probabilistic classification (Marquand et al., 2010; Rasmussen & Williams, 2006; Rocha-Rego et al., 2014).

In a clinical scenario, where diseases are heterogeneous among individuals, GPC has the advantage of include predictive uncertainty in the results, instead of a pure categorical classification, such as SVM (Mourão-Miranda et al., 2012). Thus, probabilistic classification may be in consonance with a dimensional view of mental illness, and also quantify how close are the individuals of the studied classes. On the other hand, SVM models tend to minimize the risk of overfitting, thus generalizing well to unseen samples (Christopher, 1998). Other methods are not discussed in this paper because they weren't applied to the included studies.

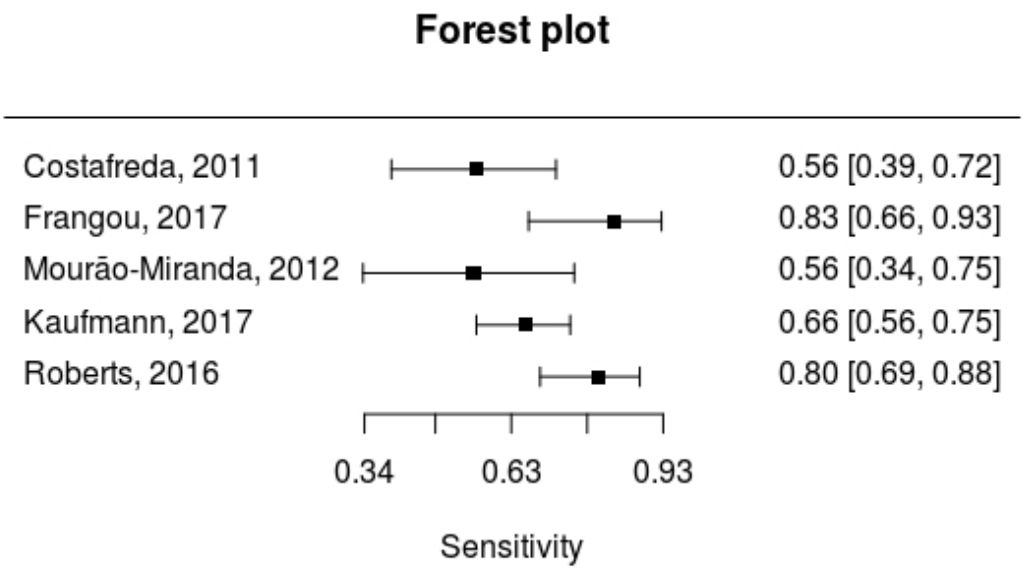
A number of limitations, however, must be taken into account when dealing with machine learning algorithms. For example, the class imbalance problem occurs when a large majority of the examples belong to a single class, which can compromise the model performance. Any model trained using an imbalanced data set may, therefore, assign new observations to the majority class (Lantz, 2015; Passos et al., 2016). This could be corrected using two techniques. In random oversampling, the examples of the minority class are replicated and added to the data set, until an adequate balance is obtained. As data of the minority class is replicated, we incur in the risk of overfitting. Undersampling, on the other hand, removes data until a sample of the majority class is balanced with the minority one. To avoid loss of information, we can build multiple models by taking random samples of the majority class and comparing them to the minority class, thus generating multiple classifiers that could be later combined (Haibo He & Garcia, 2009). Another problem occurs when the model performs well in the training data set but has a poor performance in the testing data set. This may occur potentially due to overfitting, when the model does not generalize well to previously unseen subjects data (Gollapudi, 2016; Lantz, 2015). Solutions for these problems may depend on the algorithm used and properties of the data.

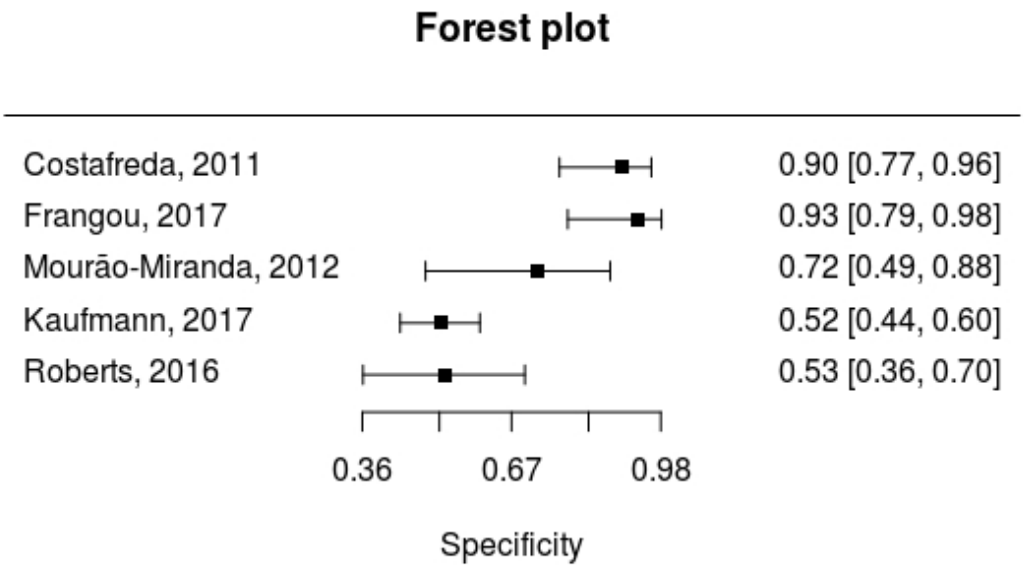
Figures S3. *Descriptive statistics for sensitivity and specificity of machine learning studies that used structural neuroimaging to differentiate bipolar patients from healthy controls*



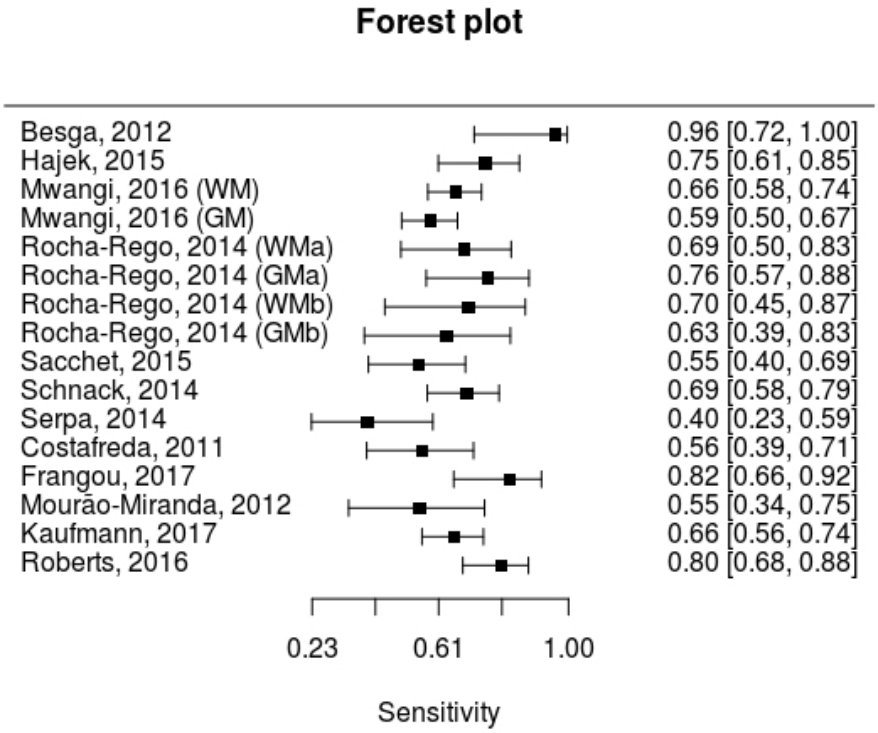


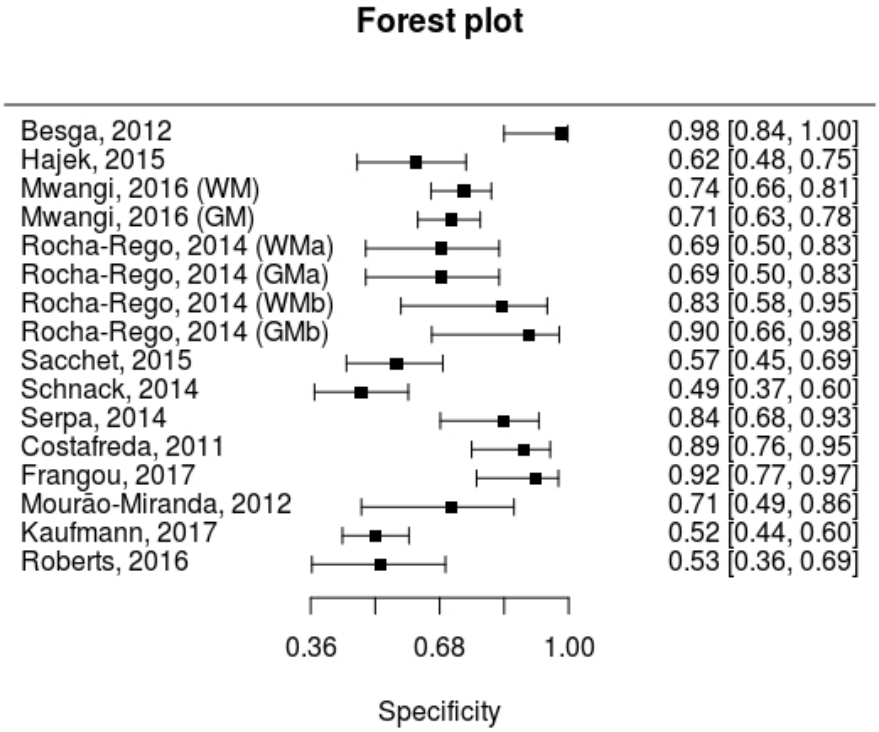
Figures S4. *Descriptive statistics for sensitivity and specificity of machine learning studies that used functional neuroimaging to differentiate bipolar patients from healthy controls*





Figures S5. *Descriptive statistics for sensitivity and specificity of machine learning studies that used both structural and functional neuroimaging to differentiate bipolar patients from healthy controls*





Bibliography

Bishop, M. C. (2006). *Pattern Recognition and Machine Learning*. Springer. Retrieved from papers3://publication/uuid/576DFF89-1DD1-434D-8337-6954496F57C4

Cao, B., Mwangi, B., Hasan, K. M., Selvaraj, S., Zeni, C. P., Zunta-Soares, G. B., & Soares, J. C. (2015). Development and validation of a brain maturation index using longitudinal neuroanatomical scans. *NeuroImage*, *117*, 311–318. https://doi.org/10.1016/j.neuroimage.2015.05.071

Cao, B., Stanley, J. A., Selvaraj, S., Mwangi, B., Passos, I. C., Zunta-Soares, G. B., & Soares, J. C. (2016). Evidence of altered membrane phospholipid metabolism in the anterior cingulate cortex and striatum of patients with bipolar disorder I: A multi-voxel 1H MRS study. *Journal of Psychiatric Research*, *81*, 48–55. https://doi.org/10.1016/j.jpsychires.2016.06.006

Christopher, J. C. B. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Fung, G., Deng, Y., Zhao, Q., Li, Z., Qu, M., Li, K., … Chan, R. C. K. (2015). Distinguishing bipolar and major depressive disorders by brain structural morphometry: a pilot study. *BMC Psychiatry*, *15*(1), 298. https://doi.org/10.1186/s12888-015-0685-5

Gollapudi, S. (2016). *Practical Machine Learning*. (Intergovernmental Panel on Climate Change, Ed.), *Packt Publishing*. Cambridge: Cambridge University Press. https://doi.org/10.1088/1751-8113/44/8/085201

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hall, M.-H., Smoller, J. W., Cook, N. R., Schulze, K., Hyoun Lee, P., Taylor, G., … Levy, D. L. (2012). Patterns of deficits in brain function in bipolar disorder and schizophrenia: A cluster analytic study. *Psychiatry Research*, *200*(2–3), 272–280. https://doi.org/10.1016/j.psychres.2012.07.052

Huys, Q. J. M., Maia, T. V, & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Lantz, B. (2015). *Machine Learning with R - Second Edition*. (Intergovernmental Panel on Climate Change, Ed.), *Packt Publishing*. Cambridge: Cambridge University Press. Retrieved from http://books.google.com/books?id=ZQu8AQAAQBAJ&printsec=frontcover&dq=intitle:Machine+Learning+with+R&hl=&cd=1&source=gbs_api%5Cnpapers2://publication/uuid/46164A51-A282-4F67-8397-9FA79F39B5B7

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., & Mourão-Miranda, J.

(2010). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage*, *49*(3), 2178–2189. https://doi.org/10.1016/j.neuroimage.2009.10.072

Mourão-Miranda, J., Oliveira, L., Ladouceur, C. D., Marquand, A., Brammer, M., Birmaher, B., … Phillips, M. L. (2012). Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0029482

Mwangi, B., Soares, J. C., & Hasan, K. M. (2014). Visualization and unsupervised predictive clustering of high-dimensional multimodal neuroimaging data. *Journal of Neuroscience Methods*, *236*, 19–25. https://doi.org/10.1016/j.jneumeth.2014.08.001

Passos, I. C., Mwangi, B., Cao, B., Hamilton, J. E., Wu, M.-J., Zhang, X. Y., … Soares, J. C. (2016). Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach. *Journal of Affective Disorders*, *193*, 109–116. https://doi.org/10.1016/j.jad.2015.12.066

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning.* MIT Press. https://doi.org/10.1142/S0129065704001899

Rocha-Rego, V., Jogia, J., Marquand, A. F., Mourao-Miranda, J., Simmons, A., & Frangou, S. (2014). Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach. *Psychological Medicine*, *44*(3), 519–532. https://doi.org/10.1017/S0033291713001013

Struyf, J., Dobrin, S., & Page, D. (2008). Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. *BMC Genomics*, *9*(1), 531. https://doi.org/10.1186/1471-2164-9-531

Table S1 – Machine learning studies predicting clinical outcomes of bipolar disorder

First author, year	Data utilized	Sample size and diagnosis[1]	Machine learning model	Accuracy	Other measures	Commentary
			DEPRESSION RELAPSE			
Salvini et al. (2015)	Demographic and clinical features.	139 BD patients	ILP RLS Aleph	Relapse Group: 85%; Non-Relapse Group: 91%.	Sensitivity: RG: 92% NRG: 73% Specificity: RG: 59% NRG: 95%	------
			MOOD CHANGES			
Faurholt-Jepsen et al. (2016)	Voice features and smartphone data; self-monitored data on illness activity; HAMD 17-item; YMRS.	28 patients with BD according to ICD-10 were included	RF	User dependent-model: 70% (depression) 61% (manic/mixed) User independent model: 68% (depression)	AUCs: User dependent-model Sensitivity: 64% (depression) Specificity: 75% (depression) Sensitivity: 71% (manic/mixed)	Only 13 patients had enough data on voice features to train a model; data shown here only for the voice features models

				74% (manic/mixed)	Specificity: 50% (manic/mixed) User independent model: Sensitivity 81% (depression) Specificity 56% (depression) Sensitivity: 97% (manic/mixed) Specificity: 52% (manic/mixed)	
Gentili et al. (2017)	Heart rate variability (HRV) features acquired by a wearable monitoring system.	8 BD patients	SVM	Independent classification: 68.57% (SD=8.01) Normalization with previous mood state: 85.26% (SD=13.55) Normalization with random mood states: 79.30% (SD=7.10) Normalization with previous and subsequent mood state: 99.25% (SD=1.00)	-----	Models were built for each mode state alone, previous mood state, previous and subsequent mood states, and random mood states.
Valenza et al. (2013)	Biosignals via ECG, RR interval and respiration features.	3 euthymic BD patients	MLP PCA LDC QDC MOG	88-97%	-----	Model depends on initial clinical observations; small sample.

			k-NN KSOM,			
Valenza et al. (2014)	ECG, respirogram and body posture data.	8 BD patients (depression, mixed state, euthymia and hypomania)	SVM	70.8 to 96.25% to differentiate mood states	-----	-----
			SUICIDE			
Levey et al. (2016)	50 validated biomarkers coupled with CFI-S and SASS scores.	Discovery cohort: 12 Validation cohort: 6 Test cohort for SI: 33 Test cohort for future hospitalizations for suicidality: 24 BD, MDD, SA and SCZ patients	Not specified	NA	AUCs (all disorders): 0.82 for SI (p-value 0.003) 0.78 for future hospitalizations (p-value 0.032)	Included only female participants; analysis for individual biomarkers not shown here
Niculescu et al. (2015)	Microarray gene expression studies using a CFG approach; CFI-S scores; SASS scores.	Discovery cohort: 37 Validation cohort: 26 Test cohort for SI: 108 Test cohort for future hospitalizations for suicidality: 157	Not specified	NA	AUCs (BD subjects): 0.98 for SI (p-value 1.19e-6) 0.94 for future hospitalizations for suicidality (p-value 0.0021)	All participants were male; analysis for individual biomarkers not shown here

		BD, MDD, SA and SCZ patients				
Passos et al. (2016)	Data from studies that reported clinical and demographic risk factors for suicide.	144 subjects: - 114 with BD; - 30 with MDD.	RVM SVM LASSO	RVM: 72%; SVM: 64.7%; LASSO: 68%.	Sensitivity RVM: 72.1%; SVM: 58.1%; LASSO: 55.8%. Specificity: RVM: 71.3%; SVM: 71.3%; LASSO: 80.2%. AUC: RVM: 0.77; SVM: 0.65; LASSO: 0.73.	-----

TREATMENT RESPONSE AND ADVERSE EFFECTS

Castro et al. (2016)	Electronic health records (EHRs) of patients after (or not) lithium treatment.	5751 subjects: - 1445 with RI; -4306 HC.	Logistic regression	NA	AUC: 0.81 in training and testing sets Sensitivity: 45% Specificity: 92%	-----
Wade et al. (2016)	Pre-treatment sMRI morphometric measures of caudate, putamen, pallidum and nucleus accumbens; HAM-D, MADRS, QIDS-SR scores.	86 subjects: - 45 UD - 8 BD (depressive episode only) - 33 HC	SVM	89% for combined baseline morphometric measures and combined mood scale	AUC: 0.90	Small sample of BD patients

Table S2 – Machine learning studies using unsupervised or semi-supervised algorithms on bipolar disorder

First author, year	Data utilized	Sample size and diagnosis[1]	Machine learning model	Clusters found	Commentary
Bansal et al. (2012)	TI-weighted MR images from the cortex, amygdala, and hippocampus.	-42 Healthy Children; -40 Healthy Adults; -71 TS children; -36 TS adults; -41 ADHD children; -26 BD adults; -65 adults SCZ; -66 High risk for MDD (12C, 54A); -65 Low risk for MDD (31C, 34A).	Hierarchical Clustering	Semi-supervised learning – authors classified psychiatric disorders based on the morphological features of cortical and subcortical brain regions. Models clustered brains in two and four clusters prior to the classification task.	Classification accuracy: 100% (BD vs HC). 99.99% (BD vs SCZ). 96.4% (BD vs HC). 100% (BD vs SCZ).
Hall et al. (2012)	EEG neurophysiological profiles.	469 subjects: - 49 with non-psychotic BD; - 68 with BD; - 19 non-psychotic SCZ; - 59 with SCZ; - 14 non-psychotic SA; - 20 SA;	k-means clustering	Found three clusters, named by the authors as “globally impaired”, “sensory processing” and “high cognitive”. Results does not support the distinction between BD and SCZ proposed in the DSM.	----

		- 230 HC.			
Wahlund et al. (1998)	Clinical symptoms with CPRS (Comprehensive Psychopathologial Rating Scale) and MAO activity.	28 subjects: - 28 with UD.	PCA	Found four clusters of patients, cluster III had 3 bipolar patients, cluster IV one, and one bipolar patients was an outlier, but closer to clusters III and IV than I and II.	----
Wu et al. (2016b)	Neurocognitive data obtained from CANTAB and neuroimage data.	70 subjects: -70 with BD type I, II or NOS	k-means clustering PCA LASSO Elastic Net	Phenotype I and II, with accuracies of 57% and 92%, respectively, when distinguishing this profiles from healthy controls.	Label prediction after clustering: LASSO: 94% accuracy (I vs. II) Elastic Net: 75.9% (I vs. II) 92%. (II vs. HC)

Abbreviations:

ADHD, attention deficit hiperactivity disorder; Aleph, A learning engine for proposing hypotheses; BD, bipolar disorder; BDd, bipolar disorder depressed; CANTAB, Cambridge Neurocognitive Test Automated Battery; CFG, Convergent Functional Genomics; CFI-S, Convergent Functional Information for Suicidality; DSM, Diagnostic and Statistical Manual of Mental Disorders; HAMD, Hamilton Depression Rating Scale; HC, healthy controls; ILP, Inductive logic programming; k-NN, k-Nearest neighbor; LASSO, Least Absolute Shrinkage and Selectin Operator; LDC, Linguistic data consortium; MADRS, Montgomery-Åsberg Depression Rating Scale; MDD, major depressive disorder; MLP, Multilayer Perceptron; MoG, Mixture of Gaussian; NOS, not otherwise specified; PCA, Principal Component Analysis; QDC, Quadratic classifier; QIDS-SR, Quick Inventory of Depressive Symptomatology Self Report; RF, Random Forest; RI, renal insufficiency; RLS, Relational learning systems; RVM, Relevance Vector Machine; SA, schizoaffective disorder; SASS, Simplified Affective State Scale; SCZ, schizophrenia; SVM, Support Vector Machine; TS, Tourette syndrome; UD, unipolar disorder; YMRS, Young Mania Rating Scale.

[1]All studies used DSM-IV criteria for diagnosis, except when specified otherwise. Valenza et al, 2013, didn't specify diagnostic criteria.

6.2. Artigo 2

Título: **"Machine learning guided intervention trials in mental health: a systematic review and methodological recommendations"**

Atualmente sendo reescrito conforme sugestão dos revisores da *Molecular Psychiatry* para nova submissão nesta revista.

Fator de impacto da revista: 11.973

Gmail

Diego Librenza Garcia <librenzagarcia@gmail.com>

2019MP000094 Receipt of New Paper by Molecular Psychiatry

1 message

MolecularPsychiatry@us.nature.com <MolecularPsychiatry@us.nature.com> 28 January 2019 at 05:39
Reply-To: MolecularPsychiatry@us.nature.com
To: librenzagarcia@gmail.com

Dear Dr Librenza-Garcia,

Please note that you are listed as a co-author on the manuscript "Machine learning guided intervention trials in mental health: a systematic review and methodological recommendations" (reference number: 2019MP000094), which was recently submitted to Molecular Psychiatry.

The corresponding author is solely responsible for communicating with the journal and managing communication between co-authors. Please contact the corresponding author directly with any queries you may have related to this manuscript.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals. Please check your account regularly and ensure that we have your current contact information.

In addition, NPG encourages all authors and reviewers to associate an Open Researcher and Contributor Identifier (ORCID) to their account. ORCID is a community-based initiative that provides an open, non-proprietary and transparent registry of unique identifiers to help disambiguate research contributions.

Access your account

Many thanks,
NPG Applications Helpdesk
Springer Nature Limited

This email has been sent through the Springer Nature Manuscript Tracking System NY-610A- Springer Nature&MTS

Confidentiality Statement:

This e-mail is confidential and subject to copyright. Any unauthorised use or disclosure of its contents is prohibited. If you have received this email in error please notify our Manuscript Tracking System Helpdesk team at http://platformsupport.nature.com .

Details of the confidentiality and pre-publicity policy may be found here http://www.nature.com/authors/policies/confidentiality.html

Privacy Policy | Update Profile

Machine learning guided intervention trials in mental health: a systematic review and methodological recommendations

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

1. Post-Graduation Program in Psychiatry and Behavioral Sciences, Federal University of Rio Grande do Sul, Porto Alegre, Brazil;

2. Laboratory of Molecular Psychiatry, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil;

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

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

5. Graduation Program in Computer Science, School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil;

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:

Diego Librenza-Garcia: librenzagarcia@gmail.com

Bruno Jaskulski Kotzian: brunokotzian@hotmail.com

Devon Watts: wattsd@mcmaster.ca

Jessica Yang: jessica.yang10@gmail.com

Pedro Ballester: pedballester@gmail.com

Benson Mwangi: benson.mwangi@gmail.com

Flávio Kapczinski: flavio.kapczinski@gmail.com

Ives Cavalcante Passos: ivescp1@gmail.com

Abstract

Clinical trials and meta-analyses in psychiatry yield group-level results, and usually do not adequately model the heterogeneity and multimorbidity observed in patients. There is a critical need for individualized predictive tools to promote personalized interventions in mental health. The aim of the present study was to systematically review intervention studies coupled with machine learning techniques to build personalized predictive models of treatment outcome. We systematically reviewed articles in PubMed, Web of Science and Embase published in any language up to July 2017. Articles were included if they assessed interventions in patients with psychiatric disorders coupled with machine learning techniques to develop predictive tools of treatment outcome. We included 61 studies that assessed predictive models of pharmacological and non-pharmacological interventions among psychiatric patients with any diagnosis (n = 47,957). Multiple types of data, including clinical assessments, serum biomarkers, electroencephalography (EEG) and neuroimaging, such as functional magnetic resonance imaging (fMRI), were used to build the predictive models. Multimodal data and more complex measures, such as EEG and fMRI, are the most promising predictors for personalized prediction of outcomes, which showed consistently high and reliable performances. Clinical data seems to yield heterogeneous results, and the addition of clinical data to more complex models do not consistently improve, and may even lower, algorithmic performance. Nevertheless, several methodological issues and lack of external validation still hinder the use of these models in clinical settings, and the scarcity of transparent methods make it difficult to assess how reliable these models are. Predictive models using machine

learning techniques can address some of the gaps of evidence-based medicine, shifting the focus from group-level results to individualized models. It is also a potential way to contemplate the heterogeneity and multi-morbidity profile of real clinical samples. Although promising, external validation and studies with larger sample sizes should be conducted to test the applicability of these techniques in clinical scenarios. Additionally, methodological challenges, such as the absence of a uniform pipeline for machine learning studies and lack of interpretability, need to be addressed.

Keywords

machine learning; big data; big data analytics; computational psychiatry; evidence-based medicine; psychiatric disorders; treatment response; side effects; predictive psychiatry; predictive analysis; pattern recognition

Introduction

Evidence-based medicine (EBM) prompted a revolution in patient assessment and prevention strategies since its introduction in research and clinical practice. Indeed, EBM lead to several improvements in methodological research standards, as well as clinical guidelines and knowledge translation [1]. The gold-standard of EBM are meta-analyses and randomized clinical trials, which assess the average group response to a given intervention [2,3]. Considering that individual patients may deviate from the average group response, it can be expected that interventions demonstrating a high degree of efficacy in clinical trials, such as lithium in bipolar disorder, may not work for a subgroup of patients [4]. Additionally, EBM cannot map the complexity of multimorbidities that are often seen in real patients, and as a result, it is unable to render tailor-made recommendations [5]. In fact, the very idiosyncrasies that characterize most patients, such as multi-morbidity profiles, are often exclusion criteria in clinical trials. Clinical trials are then performed in very controlled populations, and as such, it is difficult to generalize these findings to patients that have additional diagnoses [6,7].

It is also important to mention that statistically significant results do not necessarily translate into clinical benefit, and a frequently overlooked problem in psychiatry is the potential harm of continually prescribing marginally effective interventions. For instance, approximately 60% of patients with major depressive disorder (MDD) who receive a US-FDA-approved antidepressant fail to achieve clinical remission [8]. Similarly, approximately 30% of patients with schizophrenia do not respond to available

treatment options [9,10]. It is also known that non-response occurs in 40% of patients with bipolar depression after eight weeks of treatment with quetiapine [11]. Other first-line treatments, such as lithium, lamotrigine, olanzapine, or olanzapine-fluoxetine combination, have similar or even less favorable outcome [11–14]. Also, the addition of antidepressants to an ongoing treatment with mood stabilizers will be helpful in only a quarter of patients with bipolar depression [15]. As such, these major limitations in EBM must be addressed in order to improve the quality of care.

A potential solution to these issues involves big data analytics and machine learning techniques [16]. Big data is defined as large *volumes* of data, created at high *velocity*, in a wide *variety* of types [17]. Considering this complexity, traditional statistical tools are insufficient to analyze these massive datasets [18]. Machine learning is a subfield of artificial intelligence focused on algorithms that can extract relevant information from complex datasets [19]. Such algorithms model patterns that can be used to create individualized predictive models with different data modalities, such as neuroimaging, genetics and clinical features [20]. Furthermore, by theoretically being able to model complex functions, machines can find complex nonlinear patterns that can relate predictors to their expected outcome. Incorporating these techniques into clinical trials and observational studies will aid in the development of personalized psychiatry, by enabling more precise interventions that include patient's idiosyncrasies [21].

In the present study, we aimed to systematically review studies that used machine learning techniques to predict treatment response or side effects in patients with psychiatric disorders.

Methods

Search strategy

Three electronic databases (PubMed, Embase, and Web of Science) were examined for articles published between January 1960 and July 2017. To identify relevant studies, the following structure for the search terms was used: ("Big data analytics" and more frequently used machine learning algorithms) AND (intervention) AND (psychiatric disorders). 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 [22]. 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 and preclinical trials were excluded.

Data collection and extraction

First, the potential articles were independently screened for title and abstract contents by two researchers (DLG and BJK). 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). 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 methods.

Results

We found 6462 potential abstracts and included 61 articles in this review, six included after reference screening (figure 1). A list of the included studies as well as their most relevant characteristics and findings are detailed in table 1. Quality assessment can be seen in table 2.

Of the included studies, 29 articles used social-demographic data and/or neuropsychological tests [23–51], 4 studies used serum biomarkers [52–55], 8 studies used electroencephalographic measures[56–63] , 11 studies used neuroimaging [64–74] and 9 studies used multimodal data [75–83].

Studies using clinical, sociodemographic and neuropsychological data

There were 29 studies that used clinical and sociodemographic data, and/or neurocognitive assessments as predictors. From these studies 14 assessed mood disorders, five assessed substance use disorders (SUD), three assessed attention deficit hyperactivity disorder (ADHD), two assessed psychotic disorders, two assessed eating disorders, two assessed multiple psychiatric disorders, and one assessed patients with obsessive compulsive disorder (OCD).

Studies assessing mood disorders

Most of the studies evaluated treatment response to selective serotonin reuptake inhibitors (SSRI), including fluoxetine, sertraline, fluvoxamine and escitalopram. From five studies assessing this class of medication, the best performance was obtained by Franchini et al. with an accuracy of 97.35% in a sample of 416 patients with MDD using an artificial neural network (ANN) algorithm [51]. Nevertheless, two recent studies with larger samples, involving a more detailed description of the methods used, obtained lower accuracies. Chekroud and colleagues had an AUC of 0.7 in a sample of 4041 nonpsychotic MDD patients treated with a 12-week course of escitalopram [45], while Winterer et al, in a large observational study with 19738 depressed patients treated with fluoxetine obtained an accuracy of 71.4% using ANN [43]. Another study with a naturalistic design assessed 1014 patients with MDD that were prescribed any antidepressant (monotherapy or co-medication) by a psychiatrist, obtained AUCs ranging from 0.63-0.72 [37]. In two studies including 116 and 145 depressed patients with either MDD or bipolar depression, Serretti and colleagues achieved accuracies of 77 and 69.17%, respectively. Both studies were open label trials of fluvoxamine [40,41].

Andreescu *et al.* used decision trees to develop two algorithms based on 0.3 and 0.7 sensitivity cut-offs. Although no overall performance was reported for these models, the authors found that marked early improvement on the 17-item Hamilton Rating Scale for Depression (HAMD-17), low baseline anxiety (first model), low baseline sleep disturbance and at least moderate early improvement (second model), were associated with a greater likelihood of response in follow-up [23].

A study with pooled data from 12 clinical trials, and a total sample size of 4987 patients, achieved AUCs of 0.6409, 0.6096 and 0.6387, in the best models, to predict response to placebo, duloxetine, and five different SSRIs [48]. Iniesta and colleagues obtained AUCs of 0.60-0.72 (escitalopram), 0.63-0.70 (nortriptyline) and 0.61-0.72 (both combined) to predict treatment response in a sample of 793 patients with MDD. The authors also developed a model to predict treatment completion, with an AUC of 0.63 for both interventions combined [27]. Similar results were obtained in another study assessing 2555 MDD patients treated with citalopram or, when not responsive, switched to a different antidepressant. The AUC for the validation stage was 0.719, which was consistent with the performance in training (0.697-0.716) and testing stages (0.693-0.712). As such, the model demonstrated adequate generalizability [33]. Etkin and colleagues used cognitive computerized batteries in 1008 subjects with MDD to predict response to escitalopram, sertraline or venlafaxine, but only the model for escitalopram was statistically significant, with an accuracy of 58%. When considering dropouts as non-responsive, the accuracy of the model increased to 67% [50].

Only two studies assessed non-pharmacological interventions. Stiles-Shields and colleagues developed a model to predict response to CBT delivered face-to-face or by phone in 325 patients with MDD, achieving 85.7% accuracy as the best classifier performance [42]. In another study, authors used neurocognitive batteries to predict response to deep brain stimulation in 20 subjects with treatment resistant depression. Using an ANN algorithm, the authors achieved an AUC of 0.93. The sample, however, had only 20 patients and lacked external validation [30]. Finally, one study predicted placebo response in a sample of 1017 patients with MDD, obtaining an AUC of 0.63 while using linear regression with feature selection [32].

Studies assessing other psychiatric disorders

Four studies assessed patients with alcohol use disorder (AUD). The interventions included: internet-based therapy and internet self-help [24]; cognitive behavioral therapy with and without acamprosate [47]; medical management combined with naloxone and/or acamprosate [25]; and inpatient psychosocial treatment including detoxification and motivational counseling [31]. The best classifier was obtained by Muller and colleagues, which achieved an AUC of 0.93 using an ANN algorithm. However, the study had a small sample (146 patients). The study with the largest sample (n=1646) achieved an AUC of 0.61 and accuracy of 60%. The only study that assessed patients with stimulant use disorder, had a sensitivity of 88.60% and specificity of 60.5% to predict response to a combined intervention of group plus individual 12-step facilitation intervention [49].

From three studies assessing patients with ADHD, 2 used methylphenidate (MPH) as an intervention [28,44], while one used [35]. Wong et al. obtained 74.3 and 76.7% accuracies, with the best models, to predict inattentiveness and hyperactivity, respectively, while Johnston and colleagues obtained 77% accuracy to predict overall treatment response. In the study that predicted response to atomoxetine, predictive positive value (PPV) ranged between 73.3% and 88.9%, while negative predictive value ranged from 46.3 to 77.5%.

Koutsouleris et al assessed 334 patients in an open-label clinical trial of 5 different atypical antipsychotics, obtaining a model to predict response after 4 weeks with 71.7% accuracy using ten selected predictors in an independent sample [29]. Ruberg et al assessed 1494 patients with schizophrenia, schizoaffective disorder or schizophreniform disorder treated with 4 different atypical antipsychotics. They obtained PPV ranging from 4 to 85% and NPV of 60 to 95%. When considering the response to each intervention, Olanzapine (PPV=81%, NPV=37%) and Ziprasidone (PPV=82%, NPV=78%) showed the best results [38]. Two studies included heterogeneous samples. Politi et al obtained 97.12% accuracy to predict response in MDD patients prescribed sertraline in a naturalistic design - with the success of the treatment being defined by the clinician [36]. Another study including 1843 patients with mood disorders, psychotic disorders, generalized anxiety disorder, childhood disorder and AUD had 91% accuracy (κ=0.59) to predict treatment noncompliance [46].

One study with a small sample of patients with anorexia nervosa (n=39), obtained initial accuracies ranging from 60 to 66%, but the performance of the predictive model improved to 77-83% when using feature selection [34]. A study assessing patients with bulimia in a sample of 647 patients, had 89% accuracy to predict poor response and 68% to predict treatment response, which consisted of multiple therapies in a naturalistic design[26]. Finally, one study assessed response to SSRI, either alone or associated with risperidone, and CBT in 130 patients with OCD, obtaining an accuracy of 93.3% and AUC of 0.945 using an ANN [39].

Studies using serum biomarkers and genetics

Four studies developed predictive models using serum biological markers. Gupta et al. used thirteen Single Nucleotide Polymorphisms (SNPs), as identified in previous studies, to predict Atypical Antipsychotic (AAP) response in a sample of 371 Schizophrenic patients. This interaction model had an overall accuracy of 73.6%, sensitivity of 71.2% and specificity of 76%. In a subgroup analysis, a model of three genetic markers alongside the demographic variables of gender, age of onset and duration and severity of illness, showed an accuracy of 72%, sensitivity of 66.4% and specificity of 77.6% [54]. Amminger et al. used data from a previous randomized clinical trial, comparing long-chain omega-3 fatty acid supplementation (ω-3 PUFAs) with placebo, to predict functional improvement among 81 individuals at ultra-high risk of developing psychosis. Functional improvement was defined as an increase of fifteen or more points in the Global Assessment of Functioning (GAF) scale. Capillary gas chromatography was used to examine baseline levels of phosphatidylethanolamine

composition, a phospholipid commonly found in cell membranes within the brain, and long-chain fatty acids thought to be relevant in schizophrenia. Fatty acid composition was found to predict response to ω-3 PUFAs with an overall accuracy of 86.7%, sensitivity of 86.7% and specificity of 86.7%. Moreover, fatty acid composition predicted response to placebo with an overall accuracy of 79.2%, sensitivity of 83.3% and specificity of 75.0% [52].

Hou et al. examined the moderating effects of genetic variations on treatment response among 251 individuals with alcohol dependence. Retrospectively collected blood samples were analyzed for 21 genetic polymorphisms and three models were used as data mining techniques to identify treatment response to ondansetron. However, the accuracy of these models, and validation methods used, were not reported [55]. Guilloux et al. assessed whether baseline genetic expression could predict remission and non-remission after treatment with citalopram. As such, they measured large-scale blood transcriptome changes in 34 patients with MDD using data from an ongoing 12-week citalopram and psychotherapy trial. Clinical and transcriptome data from a previously published study was used as a validation cohort. The average cross-validation accuracy between models was 79.4%. A 13-gene model showed a non-corrected predictive value of 88%, while a 6-gene model achieved an accuracy of 76.2%, within the validation cohort. Interestingly, a model involving two genes (IFITM3 and TIMP1) and one clinical feature (QIDS score) showed an accuracy of 97%, after correcting for model-selection bias [53]. Conversely, clinical variables alone showed a

corrected accuracy of 70.6%. Of note, baseline HRSD-17 scores were a relatively poor predictor of non-remission, showing an accuracy of 57.1% within the validation cohort.

Studies using electroencephalographic measures

Three studies used EEG data to predict response to SSRI medication in MDD. A study examined whether EEG measurements, in eyes-open and eyes-closed conditions, could predict improvement in HAMD-17 scores among 22 MDD patients treated with SSRIs for 6 weeks. The authors reported an accuracy of 86.60%, sensitivity of 87.50% and specificity of 85.70% [60]. In a second study, the same authors predicted response to SSRI treatment, using pre-treatment EEG data among 22 treatment-resistant MDD patients, with an accuracy of 87.90%, sensitivity of 94.86%, and 80.93% specificity [61].

In another study, logistic regression was used to create a model of the multivariate relationship between EEG-based features and clinical outcomes. Feature extraction was performed using Wavelet-based technique (WT) and compared against both empirical mode decomposition (EMD), and short-time Fourier transform (STFT) methods. Furthermore, rank-based feature selection using relevant class labels of responders, versus nonresponders, and patients versus healthy controls, was compared to the minimum redundancy and maximum relevance (mRMR) method. A model comprising all decomposition methods, in a single feature space, showed the highest classification efficiency with 91.6 % accuracy in distinguishing responders from non-responders, and 90.5% accuracy in distinguishing MDD patients from healthy controls, respectively. The Wavelet-based technique (WT) feature extraction method

outperformed empirical mode decomposition (EMD) and short-time Fourier transform (STFT) methods. WT was able to predict responders from non-responders with 87.5% accuracy, and MDD from controls with 89.6% accuracy [63].

Two studies predicted treatment response to repetitive Transcranial Magnetic Stimulation (rTMS) using EEG data. Arns et al. used linear and non-linear EEG analyses to predict response to rTMS and concurrent psychotherapy in a sample of 90 patients with MDD. While the authors did not report the accuracy of these models, a combination of linear and nonlinear EEG achieved the AUC of 0.835 [58]. Erguzel et al. used frontal quantitative EEG (QEEG) to assess response to adjunctive rTMS in 55 subjects with treatment-resistant MDD. The neural network-based classifier optimally identified responders vs. non-responders using a 6-fold cross validation method with 89.09% accuracy, specificity of 0.909 and sensitivity of 0.9333 [59].

Two studies predicted treatment response to transcortical Direct Current Stimulation (tDCS), using EEG data. The first study by Al-Kaysi et al. used three EEG sessions, corresponding to before treatment, during active tDCS, and during sham tDCS, respectively, to predict antidepressant response in 10 MDD patients. However, the accuracy, sensitivity and specificity of the model were not reported [56]. Similarly, the second study by Al-Kaysi et al. used baseline resting state EEG data to identify treatment responders, according to improvements in mood and cognitive scores to tDCS in 10 MDD patients but did not report the accuracy of their model [57].

One study used pretreatment EEG data to retrospectively predict clozapine response among 23 patients with chronic schizophrenia. Response to clozapine was defined as ≥25% improvement between pre and post Quantitative Clinical Assessment (QCA). The authors predicted responders and non-responders, in an independent dataset, with an average accuracy of 87.12% and 89.7%, respectively [62].

Studies using neuroimaging

Three studies developed predictive models using sMRI data. Gong et al. used gray matter (GM) and white matter (WM) regions to predict antidepressant treatment response among drug-naive patients with MDD. GM and WM had 69.57% and 65.22% accuracies, respectively, but combining both methods did not improve performance [67]. GM and WM performed better in another study which also assessed antidepressant response, with accuracies ranging between 77.1-82.9% and 65.7-82.9%, respectively [69]. The latter study, however, included patients with treatment-resistant depression. Redlich and colleagues used SVM and GPC algorithms to predict ECT response using whole-brain sMRI. They included 23 patients with acute MDD treated thrice weekly for an average of 9-12 sessions. SVM and GPC achieved, 78.3 and 73.9% accuracies, respectively. Of note, both methods had 100% sensitivity, but low specificity (SVM = 50%, GPC = 40%) [72].

From eight studies using fMRI data, three used resting-state fMRI (rs-fMRI). Specifically, rs-fMRI and diffusion-weighted imaging (DWI) were used to create a predictive model in 26 patients with Social Anxiety Disorder treated with 12 weekly

sessions of CBT. Authors obtained 81% accuracy and 84 and 78% sensitivity and specificity, respectively [70]. Instead of using a classification procedure, two studies assessed treatment response among patients with MDD using machine learning regression models. Sikora et al. attempted to predict placebo response and obtained a 0.41 correlation coefficient (p=0.18) [73], while Qin and colleagues assessed patients successfully treated with either SNRI or SSRI and reported a correlation coefficient of only -0.19 [71].

Most fMRI studies used data collected during cognitive task execution. For instance, Sunderman and colleagues used an interoception task to predict treatment response to exposure-based CBT among patients with panic disorder and agoraphobia. Authors used feature selection with t-test and SVM-RFE but obtained models with low performance (38-54.2% accuracy) [74]. Similarly, Hahn et al. assessed patients with panic disorder and agoraphobia using whole-brain fMRI during a differential fear-conditioning task and was able to predict response with accuracies ranging from 73 to 82%, with 64-92% sensitivity and 67-83% specificity. The best results came from combining acquisition and extinction phases [68]. Using BOLD signal responses during self-referential criticism, Mansson et al. developed a series of models to predict response to internet-delivered CBT and Attention Bias Modification. A model using data from the anterior cingulate cortex outperformed those using other regions, achieving an AUC of 0.91. Of note, all other regions had AUCs lower than 0.5 [70]. Using a facial expression task involving depictions of sad faces, a study with 16 medication-free MDD patients attempted to predict response to CBT. Authors used PCA for

dimensionality reduction and created the model with an SVM algorithm, obtaining a sensitivity of 71% and specificity of 85% for the lowest and higher intensities [64]. Only one study used fMRI data acquired during a task performance to predict pharmacotherapy. Fleck and colleagues used fMRI and proton resonance spectroscopy coupled with a continuous performance task with emotional and neutral distractors to predict lithium response in 20 BD type I patients with first-episode mania. The authors obtained 100% accuracy in predicting treatment response and 89.8% accuracy in predicting symptom reduction [65].

Studies assessing multimodal data

There were nine studies combining two or more levels of data. Ball and colleagues used self-report clinical measures coupled with an fMRI emotion regulation task to predict CBT response among 48 patients with generalized anxiety disorder or panic disorder. Interestingly, fMRI alone showed the highest accuracy, while adding clinical data lowered the performance of the model to 73%. Clinical data, alone, showed 69% accuracy in predicting response [75]. Another study used clinical assessments with fMRI during reward processing to predict response to a 28-day program treatment for patients with primary dependence on MPH. Clinical data showed 74%, fMRI showed 72% accuracy, while the combination of both achieved 75% accuracy, with 75% sensitivity and 81% specificity [76]. On the other hand, a study predicting response to a 12-week contingency management intervention in cocaine-dependent subjects showed decreased accuracy when assessing both clinical features and PET-scan data (from 82 to 77% in the best models). When authors used an indirect measure of

motivation (cumulative clinic attendance), in conjunction with neuroimaging, a 96% accuracy was achieved at week 3, with an area under the curve (AUC) of 0.98 [81]. Patel and colleagues achieved 89.47% accuracy, 88.89% sensitivity and 90% specificity to predict treatment response in patients with late-life depression using either duloxetine, venlafaxine, nimodipine, or escitalopram. The best model involved data from the Mini Mental State Examination, rs-fMRI and sMRI, but removed all other clinical and sociodemographic variables, which were not relevant for the predictive analysis [82]. From all included studies that used multimodal data, Kim et al. showcased a greater variety of levels of data. Apart from clinical and neuroimaging data, this study also included neuropsychological tests, genetic and environmental measures, such as exposure to lead and cotinine, as quantified using serum levels. Authors used four machine learning models to predict response in an 8-week open-label trial of MPH, with SVM outperforming the other models with 84.6% and AUC of 0.84 when all data available were included. Models which did not include all data levels showed degradation in the performance of the algorithms [79].

When using both clinical data and serum biomarkers (cytokines and neurotrophins) in a model, a study showed an average AUC of 0.785 and 0.710 in predicting response to two exercise dose groups in patients with MDD [83]. Additionally, using serum biomarkers coupled with clinical data, Hinton and colleagues used the LASSO algorithm to predict acamprosate response among 120 patients with AUD. They included a panel of amino acids and derivatives that were associated with treatment response in previous studies. The models tended to classify non-responders as responders (31%

specificity) and showed lower accuracies on the testing set (61.5%) than during the training phase (71.3%), most likely due to overfitting [77]. Using five clinical and five pharmacogenetic variables, Lin et al. investigated a predictive model to clozapine response in in-patients with schizophrenia. With ANN, the model had an AUC of 0.647 using clinical data alone. When authors used a model containing only genetic variables, they obtained an AUC of 0.805, but adding clinical variables only increased the model performance slightly, to 0.821 [80]. Only one study used text analysis coupled with clinical data to assess treatment response. Hoogendorn and colleagues used attributes extracted from e-mails that patients and therapist exchanged during an internet-based guided self-help intervention for SAD. These attributes consisted of basic mailing behavior, word usage, writing style, sentiment and topic of the messages. The best models achieved 0.72 and 0.71 AUCs (RF and LR, respectively), but it is important to mention that this model predicted outcomes using data created after the start of treatment since the text attributes were collected during the first half of the intervention [78].

Discussion

The present study of machine-learning based models assessed the predictive performance of several data modalities, including clinical variables, serum biomarkers, EEG, and neuroimaging. For our purposes, we considered clinical assessments (including questionnaires), sociodemographic data, and neuropsychological tests as one level of data, given that they are all collected from clinical assessments and the output is highly dependent on the patient.

Potential applications

The findings of this review illustrate how machine learning techniques can address gaps in our current approaches to clinical interventions. When a caregiver has a patient requiring an intervention, it is uncertain whether the patient will respond adequately, even if clinical trials suggest that the treatment will be effective. Therapeutic decisions, even those based on the most updated evidence, remain a trial-and-error effort. However, clinical decisions may be enhanced with more reliable and objective predictors that could tell us, prior to treatment and with reasonable accuracy, whether a patient will respond or develop severe side effects in follow-up. Machine learning models have the potential to introduce drastic changes in clinical settings by providing an objective tool to guide intervention that clinicians can rely on to make decisions.

Although more complex measures usually result in better model performance, adding complex and expensive measures can sacrifice applicability. Considering this, an adequate model should present with enough accuracy for clinical use, with accessible data that is time efficient to gather and analyze. Even if neuroimaging, EEG, or genetics are more objective and reliable markers of treatment response, they present with higher costs relative to clinical information. If a model with clinical data shows good accuracy, and adding more expensive markers doesn't improve performance, it is logical to assume that a cost benefit analysis would not support the use of such markers. For some interventions that have a higher cost, are less available or are more invasive, such as ECT or TMS, the use of neuroimaging or EEG measures prior to the treatment may help better allocate these resources to individuals who are likely to

respond, rather than subjecting patients to aggressive treatments that they will ultimately not benefit from. However, given the fundamental role of biological systems in psychiatric disorders, such as gene-environment interaction, neuroinflammation, changes in brain connectivity among others, it is unlikely that complex phenomena can be modeled with clinical and sociodemographic information alone.

Data modalities

Given that most of these studies remain in proof-of-concept phases, or lack adequate external validation, it is not clear which predictors are more reliable to prevent treatment related outcomes. Furthermore, it is possible that different interventions may benefit from different types of data. When evaluating responses to rTMS or ECT, for instance, neuroimaging and EEG measures may be better candidates, while genomic and metabolomics biomarkers may be more useful in predicting pharmacological interventions. Clinical data appears to show reasonable performance in some studies, although no study had adequate external validation in large samples. Furthermore, there are several limitations in using clinical and sociodemographic data. For instance, questionnaire scores based on clinical severity depends on a highly complex interaction between patient ability to report symptoms, clinician interpretation and an adequate patient-doctor relationship. In consequence, there is always the risk of information being lost, or patients failing to disclose relevant information. Additionally, self-report data may be difficult to access in some patients, and when available, it may be affected by memory bias, especially regarding past

symptomatology and events. Moreover, it requires a highly standardized system of measures and assessment that needs to be consistently followed in different sites.

Some of the included studies in this review illustrate this point. For instance, four studies found objective markers, such as gene expression and fMRI data, to demonstrate greater predictive accuracy, relative to clinical models. Of note, one found that an fMRI-based model outperformed both clinical and combined models. Exceptions to this occurred in two studies, which found adding clinical variables slightly improved model performance.

Methodological issues

It is a common practice in computer science research to share data and scripts used in machine learning models through platforms such as GitHub, Dryad or the Harvard Database Network. This enables other groups to reproduce results, replicate experiments and even improve analysis performance or adapt models to different scenarios. Ultimately, data sharing promotes a collective quality control of the methodology and ensure the veracity of data collected by their peers. Among studies included in this review, not one made either their data or analyses available. Several other issues can also be reported. Frequently, studies do not mention how they dealt with the class imbalance problem, how missing data was handled, or how the algorithms parameters were tuned. Such information is relevant given that they may create models with overfitting, with inflated accuracies and low ability to make predictions outside of the study context. It appears that studies in this field focus

either on the technical aspects of the algorithms, or primarily focus on the psychiatric disorders themselves, without adequately handling both dimensions simultaneously. A greater emphasis on addressing these issues is needed in the field, for the reader to understand the procedures and to assess the quality of the analysis.

The optimal algorithm by which to model treatment response remains uncertain. It is also difficult to know a priori which algorithm will be more appropriate, although some particularities of the data may guide this choice. Studies using feature selection should clarify if the selection was made in an embedded process, when both model and feature selection are made simultaneously, or if they used different samples to select and test the model. The use of the same dataset to select features and test the model result in a problem known as circular analysis (double-dipping) that usually results in inflated and biased accuracies. For studies to be reliable, they should address all these points in their methodologies and how the authors solved or circumvent these issues.

Redesigning clinical trials

Novel drug development has recently shifted from a strict focus on overall efficacy toward an interest in secondary outcomes, such as quality of life, cognitive enhancement, pain relief, and functionality, since most new drugs do not show greater efficacy than previously established therapeutics [84,85]. There is also a growing interest in establishing interventions with fewer side effects, thus avoiding the harmful long-term consequences of current pharmacological agents. This problem is illustrated by the metabolic effects of some antipsychotic drugs, which can increase co-morbities

and mortality among psychiatric patients. In order to achieve these goals, it is important to move beyond controlled trials with strict inclusion and exclusion criteria and group-level results, toward a model that can inform how each subject will respond to an intervention [86]. This is important not only in terms of mitigating an acute episode or preventing relapse, but also in avoiding patient harm. Another challenge for individual treatment is the correct dosage - the use of standardized dosages may ignore genetic and metabolic heterogeneity, and fast and ultrafast metabolizers may be incorrectly classified as non-responders, hindering algorithm performance. Analogously, slow metabolizers may respond with lower dosage and dropout prematurely because of important side effects.

None of the included studies used digital phenotyping, a moment-by-moment quantification of the interaction of an individual with an electronic device - most commonly, a smartphone [87]. The method by which patients interact with their smartphones may be relevant markers of mood, behavior and cognition. In consequence, these markers could also perhaps be used as predictors of treatment response. An advantage of this method is that it cannot only collect active user inputs, but also harvest passively collected data. This allows for a real-time analysis of the data, instead of single point assessments such as neuroimaging or EEG, that lose all information between two assessments. Regardless of the specific data assessed, it is still unclear how much data is needed to predict clinical effect of interventions at an individual-level. Nonetheless, large samples are needed, given that psychiatric

disorders are very heterogeneous and learning algorithms improve as a result of more examples.

Current challenges and critical needs to advance studies in the field

There is plenty of room for improvement in the field, given the current state of affairs in interventional studies. Clinical trials and observational studies should be designed with a greater focus on objective biomarkers that may show greater precision in predicting individual response and side effect risk. There is also a critical need to standardize relevant predictors, as well as their collection method, similarly to what was previously achieved in the ENIGMA consortium [88]. Indeed, one of the main limitations in machine learning studies thus far is the lack of continuity between data collection sites. Moving forward, standardized instruments with calibrated parameters will allow us to adequately validate predictive models in independent samples, and properly define and assess outcomes of interest. It is also important to mention that as of now, most machine learning intervention studies in psychiatry have used retrospectively collected data that was not originally designed for the purposes of predicting outcomes.

There is also a concern that some models lack interpretability, given that we are unable to know what mediating input and output in machine learning is - so-called black-box methods. This may result in an important loss of knowledge of the underlying mechanisms of psychiatric disorders. Thus, we may be limited in our ability to apply new procedures and treatments given that we do not fully understand how

these models works. However, a shift in this mentality may be required: if a model is accurate and can promote positive change in patient care, an immediate understanding of the underlying processes may not be necessary. Finally, another challenge is to move away from proof-of-concept studies and develop algorithms that can be tested in real clinical scenarios, with risk calculators and smartphone applications to predict response, supporting clinicians to select the best available intervention for each individual patient.

Conclusion

Although it is undeniable that RCTs brought significant advancements to patient care, personalized interventions remain a critical need in mental health [27]. Machine learning oriented interventions may help us move away from the "one size fits all" assumption of current trials by including patient heterogeneity in individualized models. Having objective tools that can enable early interventions and prevent unfavorable outcomes, such as suicide, can reduce the acute side effects and long-term impairments that are caused by available pharmacological agents [89]. Although promising, most studies in the field are still limited to research scenarios with low applicability and, frequently, they lack external validation in independent samples. Importantly, most studies are not dealing with big data, but are instead applying machine learning analysis to small samples. There is also need for more standardized assessments and greater transparency in the methods used, to guarantee that the created models are reliable outside of research contexts.

Despite these limitations, machine learning based studies are promising venues to explore that can incorporate individual features through the assessment of multiple levels of data, yielding integrative, applicable and personalized models to predict interventions outcomes. Moreover, as technology becomes increasingly sophisticated and ubiquitous in our daily lives, patient interactions with electronic devices represents a promising avenue for clinically relevant data to improve intervention and assessment strategies [87]. The interface between patients, psychiatrists, and devices may redefine mental health assessment, and address longstanding gaps in our knowledge and treatments [90].

Author's Contributions

Diego Librenza Garcia, Bruno Jaskulski Kotzian, Jessica Yang and Ives Cavalcante Passos participated in the literature search, writing, and in the approval of the final manuscript. Devon Watts, Pedro Ballester, Benson Mwangi and Flavio Kapczinski participated in the writing and in the approval of the final manuscript.

Conflict of interest statements

Diego Librenza Garcia, Bruno Jaskulski Kotzian, Jessica Yang, Devon Watts, Pedro Ballester, and Benson Mwangi report no biomedical financial interests or potential conflicts of interest. Ives Cavalcante Passos reports consulting fees from Torrent/Omnifarma, and previous funding from INCT - CNPq and 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 Canada Foundation for Innovation - CFI, outside the submitted work.

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. Greenhalgh, T., Howick, J. & Maskrey, N. Evidence based medicine: a movement in crisis? BMJ (2014). doi:10.1136/bmj.g3725

3. Djulbegovic, B. & Guyatt, G. H. Progress in evidence-based medicine: a quarter century on. The Lancet (2017). doi:10.1016/S0140-6736(16)31592-6

4. Harrison, P. J. et al. Innovative approaches to bipolar disorder and its treatment. Ann. N. Y. Acad. Sci. (2016). doi:10.1111/nyas.13048

5. 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

6. Pallmann, P. et al. Adaptive designs in clinical trials: Why use them, and how to run and report them. BMC Medicine (2018). doi:10.1186/s12916-018-1017-7

7. O'Hara, R., Beaudreau, S. A., Gould, C. E., Froehlich, W. & Kraemer, H. C. Handling clinical comorbidity in randomized clinical trials in psychiatry. J. Psychiatr. Res. (2017). doi:10.1016/j.jpsychires.2016.11.006

8. Aaronson, S. T. et al. A 5-year observational study of patients with treatment-resistant depression treated with vagus nerve stimulation or treatment as usual: Comparison of response, remission, and suicidality. Am. J. Psychiatry (2017). doi:10.1176/appi.ajp.2017.16010034

9. Mouchlianitis, E., McCutcheon, R. & Howes, O. D. Brain-imaging studies of treatment-resistant schizophrenia: A systematic review. The Lancet Psychiatry (2016). doi:10.1016/S2215-0366(15)00540-4

10. Siskind, D., McCartney, L., Goldschlager, R. & Kisely, S. Clozapine v. first- and second-generation antipsychotics in treatment-refractory schizophrenia: Systematic review and meta-analysis. British Journal of Psychiatry (2016). doi:10.1192/bjp.bp.115177261

11. De Fruyt, J. et al. Second generation antipsychotics in the treatment of bipolar depression: A systematic review and meta-analysis. Journal of Psychopharmacology (2012). doi:10.1177/0269881111408461

12. Geddes, J. R., Calabreso, J. R. & Goodwin, G. M. Lamotrigine for treatment of bipolar depression: Independent meta-analysis and meta-regression of individual patient data from five randomised trials. British Journal of Psychiatry (2009). doi:10.1192/bjp.bp.107.048504

13. Sidor, M. M. & MacQueen, G. M. Antidepressants for the acute treatment of bipolar depression: A systematic review and meta-analysis. Journal of Clinical Psychiatry (2011). doi:10.4088/JCP.09r05385gre

14. Taylor, D. M., Cornelius, V., Smith, L. & Young, A. H. Comparative efficacy and acceptability of drug treatments for bipolar depression: A multiple-treatments meta-analysis. Acta Psychiatr. Scand. (2014). doi:10.1111/acps.12343

15. Sachs, G. S. et al. Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. N. Engl. J. Med. (2007). doi:10.1056/NEJMoa064135

16. Scott, I. A. Machine Learning and Evidence-Based Medicine. Ann. Intern. Med. 169, 44 (2018).

17. Dinov, I. D. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. GigaScience (2016). doi:10.1186/s13742-016-0117-6

18. Viceconti, M., Hunter, P. & Hose, R. Big Data, Big Knowledge: Big Data for Personalized Healthcare. IEEE J. Biomed. Heal. Informatics (2015). doi:10.1109/JBHI.2015.2406883

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

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

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

22. 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

23. Andreescu, C. et al. Empirically derived decision trees for the treatment of late-life depression. Am. J. Psychiatry (2008). doi:10.1176/appi.ajp.2008.07081340

24. Blankers, M., Koeter, M. W. J. & Schippers, G. M. Baseline predictors of treatment outcome in Internet-based alcohol interventions: A recursive partitioning

analysis alongside a randomized trial. BMC Public Health (2013). doi:10.1186/1471-2458-13-455

25. Gueorguieva, R. et al. Predictors of Abstinence from Heavy Drinking During Treatment in COMBINE and External Validation in PREDICT. Alcohol. Clin. Exp. Res. (2014). doi:10.1111/acer.12541

26. Hannöver, W., Richard, M., Hansen, N. B., Martinovich, Z. & Kordy, H. A classification tree model for decision-making in clinical practice: An application based on the data of the german multicenter study on eating disorders, Project TR-EAT. Psychother. Res. (2002). doi:10.1080/713664470

27. Iniesta, R., Stahl, D. & McGuffin, P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychological Medicine (2016). doi:10.1017/S0033291716001367

28. Johnston, B. A., Coghill, D., Matthews, K. & Steele, J. D. Predicting methylphenidate response in attention deficit hyperactivity disorder: A preliminary study. J. Psychopharmacol. (2015). doi:10.1177/0269881114548438

29. 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 (2016). doi:10.1016/S2215-0366(16)30171-7

30. McInerney, S. J. et al. Neurocognitive Predictors of Response in Treatment Resistant Depression to Subcallosal Cingulate Gyrus Deep Brain Stimulation. Front. Hum. Neurosci. (2017). doi:10.3389/fnhum.2017.00074

31. Müller, S. E., Weijers, H. G., Böning, J. & Wiesbeck, G. A. Personality traits predict treatment outcome in alcohol-dependent patients. Neuropsychobiology (2008). doi:10.1159/000147469

32. Nelson, J. C. et al. Predictors of remission with placebo using an integrated study database from patients with major depressive disorder. Curr. Med. Res. Opin. (2012). doi:10.1185/03007995.2011.654010

33. Perlis, R. H. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol. Psychiatry (2013). doi:10.1016/j.biopsych.2012.12.007

34. Pohjolainen, V. et al. Bayesian prediction of treatment outcome in anorexia nervosa: A preliminary study. Nord. J. Psychiatry (2015). doi:10.3109/08039488.2014.962612

35. Block, S. L. et al. Post hoc analysis: Early changes in ADHD-RS items predict longer term response to atomoxetine in pediatric patients. Clin. Pediatr. (Phila). (2010). doi:10.1177/0009922810368134

36. Politi, E., Franchini, L., Spagnolo, C., Smeraldi, E. & Bellodi, L. Supporting tools in psychiatric treatment decision-making: Sertraline outcome investigation with artificial neural network method. Psychiatry Res. (2005). doi:10.1016/j.psychres.2004.07.011

37. Riedel, M. et al. Clinical predictors of response and remission in inpatients with depressive syndromes. J. Affect. Disord. (2011). doi:10.1016/j.jad.2011.04.007

38. Ruberg, S. J. et al. Identification of early changes in specific symptoms that predict longer-term response to atypical antipsychotics in the treatment of patients with schizophrenia. BMC Psychiatry (2011). doi:10.1186/1471-244X-11-23

39. Salomoni, G. et al. Artificial neural network model for the prediction of obsessive-compulsive disorder treatment response. J. Clin. Psychopharmacol. (2009). doi:10.1097/JCP.0b013e3181aba68f

40. Serretti, A. et al. Clinical prediction of antidepressant response in mood disorders: Linear multivariate vs. neural network models. Psychiatry Res. (2007). doi:10.1016/j.psychres.2006.07.009

41. Serretti, A., Zanardi, R., Mandelli, L., Smeraldi, E. & Colombo, C. A neural network model for combining clinical predictors of antidepressant response in mood disorders. J. Affect. Disord. (2007). doi:10.1016/j.jad.2006.08.008

42. Stiles-Shields, C., Corden, M. E., Kwasny, M. J., Schueller, S. M. & Mohr, D. C. Predictors of outcome for telephone and face-to-face administered cognitive behavioral therapy for depression. Psychol. Med. (2015). doi:10.1017/S0033291715001208

43. Winterer, G., Ziller, M. & Linden, M. Classification of observational data with artificial neural networks versus discriminant analysis in pharmacoepidemiological studies can outcome of fluoxetine treatment be predicted? Pharmacopsychiatry (1998). doi:10.1055/s-2007-979333

44. Wong, H. K. et al. Personalized medication response prediction for attention-deficit hyperactivity disorder: Learning in the model space vs. learning in the data space. Front. Physiol. (2017). doi:10.3389/fphys.2017.00199

45. Chekroud, A. M. et al. Cross-trial prediction of treatment outcome in depression: A machine learning approach. The Lancet Psychiatry (2016). doi:10.1016/S2215-0366(15)00471-X

46. Compton, M. T., Rudisch, B. E., Weiss, P. S., West, J. C. & Kaslow, N. J. Predictors of psychiatrist-reported treatment-compliance problems among patients in routine U.S. psychiatric care. Psychiatry Res. (2005). doi:10.1016/j.psychres.2005.07.009

47. Connor, J. P., Symons, M., Feeney, G. F. X., Young, R. M. & Wiles, J. The application of machine learning techniques as an adjunct to clinical decision making in alcohol dependence treatment. Subst. Use Misuse (2007). doi:10.1080/10826080701658125

48. S., D. et al. Application of the Gradient Boosted method in randomised clinical trials: Participant variables that contribute to depression treatment efficacy of duloxetine, SSRIs or placebo. J. Affect. Disord. (2014). doi:http://dx.doi.org/10.1016/j.jad.2014.05.014

49. Doyle, S. R. & Donovan, D. M. Applying an ensemble classification tree approach to the prediction of completion of a 12-step facilitation intervention with stimulant abusers. Psychology of Addictive Behaviors (2014). doi:10.1037/a0037235

50. Etkin, A. et al. A Cognitive-Emotional Biomarker for Predicting Remission with Antidepressant Medications: A Report from the iSPOT-D Trial. Neuropsychopharmacology (2015). doi:10.1038/npp.2014.333

51. Franchini, L. et al. A neural network approach to the outcome definition on first treatment with sertraline in a psychiatric population. Artif. Intell. Med. (2001). doi:10.1016/S0933-3657(01)00088-4

52. 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. Transl. Psychiatry (2015). doi:10.1038/tp.2014.134

53. Guilloux, J. P. et al. Testing the predictive value of peripheral gene expression for nonremission following citalopram treatment for major depression. Neuropsychopharmacology (2015). doi:10.1038/npp.2014.226

54. Gupta, M. et al. Identifying a predictive model for response to atypical antipsychotic monotherapy treatment in south Indian schizophrenia patients. Genomics (2013). doi:10.1016/j.ygeno.2013.02.002

55. Hou, J. et al. Subgroup Identification in Personalized Treatment of Alcohol Dependence. Alcohol. Clin. Exp. Res. (2015). doi:10.1111/acer.12759

56. Al-Kaysi, A. M., Al-Ani, A., Loo, C. K., Breakspear, M. & Boonstra, T. W. Predicting brain stimulation treatment outcomes of depressed patients through the classification of EEG oscillations. in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS (2016). doi:10.1109/EMBC.2016.7591915

57. Al-Kaysi, A. M. et al. Predicting tDCS treatment outcomes of patients with major depressive disorder using automated EEG classification. J. Affect. Disord. (2017). doi:10.1016/j.jad.2016.10.021

58. Arns, M., Cerquera, A., Gutiérrez, R. M., Hasselman, F. & Freund, J. A. Non-linear EEG analyses predict non-response to rTMS treatment in major depressive disorder. Clin. Neurophysiol. (2014). doi:10.1016/j.clinph.2013.11.022

59. Erguzel, T. T. et al. Neural network based response prediction of rTMS in major depressive disorder using QEEG cordance. Psychiatry Investig. (2015). doi:10.4306/pi.2015.12.1.61

60. Khodayari-Rostamabad, A., Hasey, G. M., MacCrimmon, D. J., Reilly, J. P. & Bruin, H. de. A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. Clin. Neurophysiol. (2010). doi:10.1016/j.clinph.2010.05.009

61. Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G., De Bruin, H. & MacCrimmon, D. Using pre-treatment EEG data to predict response to SSRI treatment for MDD. in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10 (2010). doi:10.1109/IEMBS.2010.5627823

62. 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. Clin. Neurophysiol. (2013). doi:10.1016/j.clinph.2013.04.010

63. Mumtaz, W., Xia, L., Yasin, M. A. M., Ali, S. S. A. & Malik, A. S. A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. PLoS One (2017). doi:10.1371/journal.pone.0171409

64. Costafreda, S. G., Khanna, A., Mourao-Miranda, J. & Fu, C. H. Y. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. Neuroreport (2009). doi:10.1097/WNR.0b013e3283294159

65. Fleck, D. E. et al. Prediction of lithium response in first-episode mania using the LITHium Intelligent Agent (LITHIA): Pilot data and proof-of-concept. Bipolar Disord. (2017). doi:10.1111/bdi.12507

66. Whitfield-Gabrieli, S. et al. Brain connectomics predict response to treatment in social anxiety disorder. Mol. Psychiatry (2016). doi:10.1038/mp.2015.109

67. Gong, Q. et al. Prognostic prediction of therapeutic response in depression using high-field MR imaging. Neuroimage (2011). doi:10.1016/j.neuroimage.2010.11.079

68. Hahn, T. et al. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiatry (2015). doi:10.1001/jamapsychiatry.2014.1741

69. F., L. et al. Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS One (2012). doi:http://dx.doi.org/10.1371/journal.pone.0040968

70. 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 (2015). doi:10.1038/tp.2015.22

71. Qin, J. et al. Predicting clinical responses in major depression using intrinsic functional connectivity. Neuroreport (2015). doi:10.1097/WNR.0000000000000407

72. Redlich, R. et al. Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. JAMA Psychiatry (2016). doi:10.1001/jamapsychiatry.2016.0316

73. Sikora, M. et al. Salience Network Functional Connectivity Predicts Placebo Effects in Major Depression. Biol. Psychiatry Cogn. Neurosci. Neuroimaging (2016). doi:10.1016/j.bpsc.2015.10.002

74. Sundermann, B. et al. Support vector machine analysis of functional magnetic resonance imaging of interoception does not reliably predict individual outcomes of cognitive behavioral therapy in panic disorder with agoraphobia. Front. Psychiatry (2017). doi:10.3389/fpsyt.2017.00099

75. Ball, T. M., Stein, M. B., Ramsawh, H. J., Campbell-Sills, L. & Paulus, M. P. Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology (2014). doi:10.1038/npp.2013.328

76. Gowin, J. L., Ball, T. M., Wittmann, M., Tapert, S. F. & Paulus, M. P. Individualized relapse prediction: Personality measures and striatal and insular activity during reward-processing robustly predict relapse. Drug Alcohol Depend. (2015). doi:10.1016/j.drugalcdep.2015.04.018

77. Hinton, D. J. et al. Metabolomics biomarkers to predict acamprosate treatment response in alcohol-dependent subjects. Sci. Rep. (2017). doi:10.1038/s41598-017-02442-4

78. Hoogendoorn, M., Berger, T., Schulz, A., Stolz, T. & Szolovits, P. Predicting Social Anxiety Treatment Outcome Based on Therapeutic Email Conversations. IEEE J. Biomed. Heal. Informatics (2017). doi:10.1109/JBHI.2016.2601123

79. Kim, J. W., Sharma, V. & Ryan, N. D. Predicting methylphenidate response in ADHD using machine learning approaches. Int. J. Neuropsychopharmacol. (2015). doi:10.1093/ijnp/pyv052

80. Lin, C. C. et al. Artificial neural network prediction of clozapine response with combined pharmacogenetic and clinical data. Comput. Methods Programs Biomed. (2008). doi:10.1016/j.cmpb.2008.02.004

81. Luo, S. X., Martinez, D., Carpenter, K. M., Slifstein, M. & Nunes, E. V. Multimodal predictive modeling of individual treatment outcome in cocaine dependence with combined neuroimaging and behavioral predictors. Drug Alcohol Depend. (2014). doi:10.1016/j.drugalcdep.2014.04.030

82. Patel, M. J. et al. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int. J. Geriatr. Psychiatry (2015). doi:10.1002/gps.4262

83. 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. Depress. Anxiety (2017). doi:10.1002/da.22670

84. Leucht, S. et al. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: A multiple-treatments meta-analysis. Lancet (2013). doi:10.1016/S0140-6736(13)60733-3

85. Yatham, L. N. et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) and International Society for Bipolar Disorders (ISBD) 2018 guidelines for the management of patients with bipolar disorder. Bipolar Disord. (2018). doi:10.1111/bdi.12609

86. 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. Mol. Psychiatry (2018). doi:10.1038/s41380-018-0250-y

87. Insel, T. R. Digital phenotyping: Technology for a new science of behavior. JAMA - Journal of the American Medical Association (2017). doi:10.1001/jama.2017.11295

88. McMahon, M. A. B. & Thompson, P. M. Enhancing neuro imaging genetics through meta analysis: global collaborations in psychiatry by the ENIGMA consortium. Eur. Neuropsychopharmacol. 27, S715 (2017).

89. Niculescu, A. B. et al. Precision medicine for suicidality: From universality to subtypes and personalization. Mol. Psychiatry (2017). doi:10.1038/mp.2017.128

90. Insel, T. R. Digital phenotyping: a global tool for psychiatry. World Psychiatry 17, 276–277 (2018).

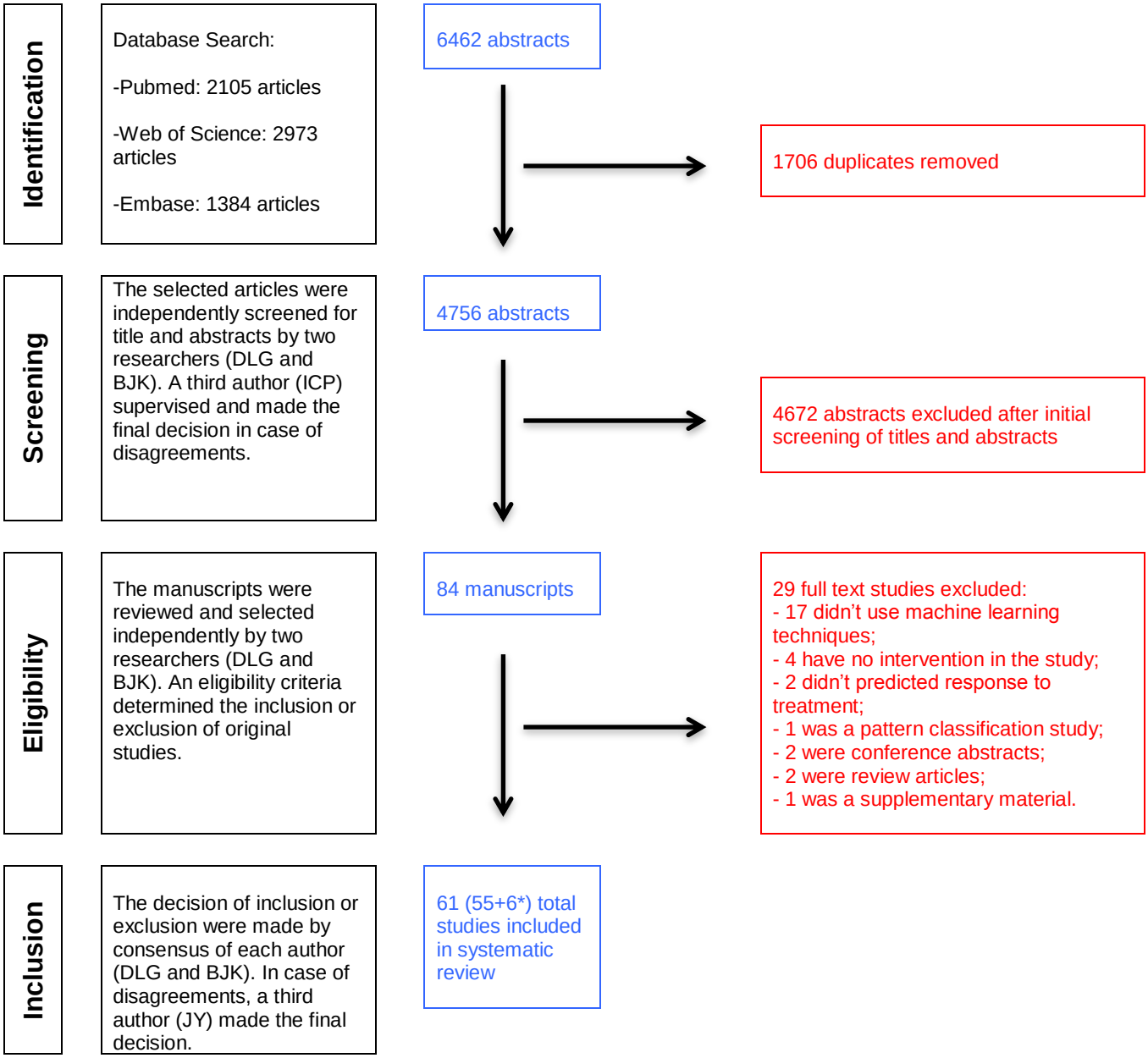


Figure 1. Flowchart of review process and study selection

*58 articles were included using the method described in this flowchart. 6 additional articles were included by searching through the references from the 58 articles included.

Table 1 – Machine learning studies predicting intervention outcomes in psychiatric disorders

First author, year	Data utilized	Sample size and diagnosis[1,2]	Intervention	Machine learning model	Accuracy	Other measures
STUDIES USING CLINICAL AND SOCIO-DEMOGRAPHIC DATA						
Andreescu, 2008	Clinical assessments from 3 different instruments	461 patients with late-life MDD (pooled data from 3 trials)	Paroxetine and nortriptyline	DT	NA	Two decision tree models created: 1) sensitivity cutoff of 0.3; and 2) sensitivity cutoff point of 0.7
Blankers, 2013	12 predictors related to alcohol use, 27 related to psychosocial functioning, and 12 sociodemographic variables	136 patients with AUD	Two internet-based interventions for AUD (Internet Therapy and Internet Self-Help	CART Feature selection: univariate regression	NA	Sensitivity: 0.34 Specificity: 0.89
Block, 2010	Individual items of the ADHD-RS at weeks 1, 2 and 3	562 patients with ADHD (5 trials)	Atomoxetine	CART	NA	PPV: 73.3-88.9% NPV: 46.3-77.5%
Chekroud, 2016	Self-report from one clinical instrument	4041 nonpsychotic MDD	12-week course of escitalopram	GBM	64.6% (internal CV) 59.6% (external validation)	AUC: 0.7
Compton, 2005	Clinical and sociodemographic predictors	1843 patients - 945 with mood disorders - 289 with SCZ or other psychotic disorder - 208 with GAD - 140 with childhood disorder - 61 with SUD	Any treatment (study predicting treatment noncompliance)	LR	91% kappa = 0.59	NA
Connor, 2007	Clinical and sociodemographic predictors	169 patients with alcohol dependence	12 week CBT versus 12 week CBT + acamprosate	DT (C5.0) BN DA	DT: 77% BN: 73% DA: 42%	NA
Dodd, 2014	14 clinical and sociodemographic predictors	4987 patients with MDD (pooled data from 12 clinical trials)	Duloxetine and five ISRS (fluoxetine, paroxetine, sertraline, escitalopram, citalopram)	tree-based GBM LR	NA	Placebo GBM: 0.6206 LR: 0.6409 Duloxetine GBM: 0.6024 LR: 0.6096 ISRS GBM: 0.6317 LR: 0.6387
Doyle, 2014	Clinical and sociodemographic	234 patients with	8-week combined group plus	CART	NA	Sensitivity: 0.886

	predictors selected from literature review	stimulant use disorder (cocaine, amphetamine or methamphetamine, and other stimulants)	individual 12-step facilitative intervention			Specificity: 0.605 PPV: 0.824 NPV: 0.719
Etkin, 2015	Computerized test battery (13 tests assessing domains of psychomotor, executive, memory-attention, processing speed, inhibition and emotional functions)	1008 subjects with first onset or recurrent nonpsychotic MDD	Escitalopram, sertraline or venlafaxine-extended release treatment (in MDD patients)	LDA	MDD vs. HC: 56% Intact vs. HC: 57% Impaired vs. HC: 91% Treatment response to Escitalopram: 58% Escitalopram and dropouts considered non-responsive: 67% HRSD-17 Remission - Intact ESC58/ SERT 51/ VEN 38 HRSD-17 Remission-Impaired - ESC65 / SERT 59/ VEN 52 HRSD-17 Response-Intact- ESC 53/ SERT 55/ VEN 38 HRSD-17 Response-Impaired ESC 65/ SERT 59/ VEN 52	HRSD-17 Remission - Impaired (Sensitivity) ESC 53/ SERT 57/ VEN 60 HRSD-17 Remission - Impaired (Specificity) ESC 59/ SERT 61/ VEN 53 HRSD-17 Response - Intact (Sensitivity) ESC 59/ SERT 60/ VEN 35 HRSD-17 Response - Intact (Specificity) ESC 46/ SERT 39/ VEN 40 HRSD-17 Response - Impaired (Sensitivity) ESC 57/ SERT 51/ VEN 57 HRSD-17 Response - Impaired (Specificity) ESC 48/ SET 48/ VEN 45 QIDS-SR16 Remission - Intact (Sensitivity) ESC 29/ SERT 57/ VEN 45 QIDS-SR16 Remission - Intact (Specificity) ESC 40/ SERT 51/ VEN 40 QIDS-SR16 Remission - Impaired (Sensitivity) ESC 79/ SERT 66/ VEN 58 QIDS-SR16 Remission - Impaired (Specificity) ESC 69/ SERT 63/ VEN 56
Franchini, 2001	Clinical and sociodemographic predictors	416 patients with MDD	Sertraline	ANN	97.35%	NA
Gueorguieva, 2015	Clinical and sociodemographic predictors	1646 subjects with alcohol dependence	Medical management, naloxone, acamprosate	CART DF	60% (external validation)	AUC: 0.61

				LR		
Hannover, 2002	Clinical and sociodemographic predictors	647 patients with eating disorder (bulimia)	Multiple interventions in a naturalistic design, including psychodynamic psychotherapy	CART	89% to predict less favourable outcomes; 68% to predict successful outcome	NA
Iniesta, 2016	Four combinations of clinical and sociodemographic data, with a total of 413 predictors	793 participants with MDD	Escitalopram and nortriptyline	ENRR	NA	AUC Overall: 0.61-0.72 Escitalopram: 0.60-0.72 Nortriptyline: 0.63-0.7 Treatment completion: 0.63
Johnston, 2014	Clinical and sociodemographic data, plus 7 neuropsychological task measures	43 patients with ADHD	12-week placebo-controlled, double blind, randomized, crossover trial of MPH	SVM-L	77%	Sensitivity: 0.54 Specificity: 0.87
Koutsoleuris, 2015	Clinical and sociodemographic data	334 patients with first-episode psychosis	Open-label clinical trial of haloperidol, amisulpride, olanzapine, quetiapine, and ziprasidone	SVM-RBF SVM-L Univariate LR Multivariate LR DT	4-week pooled: 68.5-75% 4-week LSO: 69.6-72.1% 4-week top 10: 71.7% 52-week pooled: 55.4-73.8% 52-week LSO: 67.7-72.5%	Sensitivity / Specificity 4-week pooled: 73.7-88.0% / 50.0-76.4% 4-week LSO: 65.6-70.5% / 72.7-74.5% 4-week top 10: 71.1% / 72.2% 52-week pooled: 16.7-66.7% / 73.8-95.7% 52-week LSO: 62.8-67.9% / 72.7-78.1%
McInerney, 2017	WCST (Total errors), Finger Tap Dominant Hand, and Non-Dominant Hand	20 subjects with unipolar TRD -11 responders - 9 non-responders	DBS in SCG white matter - stimulation parameters set between 3.5 and 5.0 V	ANN	AUC of 92.9% reported	The model was able to predict 8/9 non-responders and 10/11 responders correctly from the patient sample
Muller, 2008	Five personality questionnaires	146 patients with with alcohol dependence	Inpatient psychosocial treatment program, including detoxification and motivational counseling for abstinence	ANN	NA	AUC of 0.93
Nelson, 2012	Clinical and sociodemographic predictors	1017 patients with MDD	Placebo	LR with forward variable selection CART	NA	LR: 0.63 CART: 0.57
Perlis, 2013	Clinical and sociodemographic predictors	2555 patients with MDD	Sequential intervention of 1) citalopram or 2) switch to sertraline, bupropion or venlafaxine, or augmentation with bupropion or buspirone	LR NB SVM-RBF RF Feature selection with wrapper method	NA	AUC (validation): 0.719 AUC (training/testing) LR: 0.714/0.712 NB: 0.716/0.698 RF: 0.706/0.693 SVM: 0.697/0.706

Pohjolainen, 2014	Clinical and sociodemographic predictors	39 patients with AN	Routine treatment at a eating disorder clinic	NB	60-66% 77-83% (with feature selection)	NA
Politi, 2005	Clinical and sociodemographic predictors	1249 patients with any psychiatric disorder	Sertraline	ANN	Training: 96.08% Testing 97.12%	NA
Riedel, 2011	14 clinical and sociodemographic variables	1014 patients with MDD	Any antidepressant treatment prescribed by the psychiatrist, either as monotherapy or with co-medication	LR with AIC for feature selection CART	NA	AUC 0.63-0.72
Ruberg, 2011	Clinical and sociodemographic variables	1494 patients with SCZ, schizoaffective disorder or schizophreniform disorder	Olanzapine, risperidone, ziprasidone and quetiapine	CART	NA	PPV / NPV Overall: 46-85%/60-95% Olanzapine: 81%/73% Quetiapine: 40%/88% Risperidone: 66%/77% Ziprasidone: 82%/78%
Salomoni, 2009	Clinical and sociodemographic variables; three measures from neuropsychological tests	130 patients with OCD	SSRI, SSRI plus risperidone, CBT (exposure and response prevention)	LR ANN	LR: 61.5% ANN: 93,3%	AUC LR: 0.645 ANN: 0.945
Serretti, 2007	15 clinical and sociodemographic variables	116 depressed patients - 89 unipolar - 27 bipolar	6-week open-label trial with fluvoxamine	LR ANN	LR: 77% ANN: 62%	AUC LR: NA ANN: 0.769
Serretti, 2007b	Clinical and sociodemographic variables	145 depressed patients with MDD or BD	Fluvoxamine	ANN	69.17%	PPV 75.8% NPV 44.4%
Stiles-Shields, 2016	Clinical and sociodemographic variables	325 patients with MDD	CBT delivered face-to-face or by phone	CART RF	CART: 85-85.7% RF: 69.2%	NA
Winterer, 1998	Clinical and sociodemographic predictors	19738 depressed patients	Fluoxetine (observational study)	ANN DA	LDA: 71.4% ANN: 71%	NA
Wong, 2017	Clinical and sociodemographic predictors	157 patients with ADHD	Methylphenidate	Learning in the mode space SVR GPR MER	44.8- 76.7%	AUC: 0.410-0.844
First author, year	**Data utilized**	**Sample size and diagnosis[1]**	**Intervention**	**Machine learning model**	**Accuracy**	**Other measures**
STUDIES USING SERUM BIOMARKERS						
Amminger, 2015	Erythrocyte fatty acid composition of the phosphatidylethanolamine quantified via capillary gas chromatography. AA, ALA, DHA,	81 subjects at UHR of psychosis -27 males -54 females	ω-3 PUFAs vs. placebo	GPC	ω-3: 86.7% placebo: 79.6%	ω-3: sensitivity - 86.7%; specificity: 86.7% placebo: sensitivity 83.3%; specificity: 75%

	EPA, LA ,NA were also examined.					
Guilloux, 2015	Peripheral blood gene expression, including cell-specific biomarkers and promoter transcription factor enrichment	34 subjects with MDD + comorbid anxiety Independent cohort: 63 subjects with MDD, treated with citalopram, an external institution	12-weeks of citalopram and 12 sessions of psychotherapy (IPT-PS or BSP)	SVM with linear kernel	Initial cohort: 79.4% Validation cohort: 76.2%	Initial cohort: sensitivity - 66.7%; specificity - 89.5% Validation cohort: sensitivity 86.1%; specificity - 59.3%
Gupta, 2013	SNPs	371 subjects with schizophrenia -192 complete responders to AAP -179 incomplete responders to AAP	Acamprosate (one 333 mg tablet thrice daily to determine tolerance; then standard two 333 mg tablets thrice daily)	LASSO penalized selection method Sparse logistic regression based on LASSO with four-fold cross-validation	Training: 0.713 (±0.038) Testing: 0.615 (±0.029)	Training: sensitivity - 0.949 (±0.049) specificity - 0.366 (±0.174) Testing: sensitivity - 0.833 (± 0.138) specificity - 0.310 (± 0.186)
Hou, 2015	21 genetic polymorphisms	251 subjects with AUD	11-weeks of ondansetron (4μg/kg, twice daily)	IT VT	N/A	Mean difference of PHDD - IT subgroup: 17.2% - VT subgroup: 21.8% (Table 1) subgroup comparison for % of subjects with ≤1 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=0.059
First author, year	**Data utilized**	**Sample size and diagnosis[1]**	**Intervention**	**Machine learning model**	**Accuracy**	**Other measures**
STUDIES USING ELECTROENCEPHALOGRAPHIC MEASURES						
Al-Kaysi, 2016a	Three 20-minute EEG sessions	10 subjects with MDD	tDCS	SVM LDA ELM	N/A	Average error rate (3 algorithms): 0.2167 0.2 (ensemble of voters) 1/10 misclassified cognition and 3/10 misclassified mood
Al-Kaysi, 2016b	Resting state EEG at baseline - 62 single channels and the following pairs: F3, F4, C3, C4, P3, P4	10 subjects with MDD	tDCS	SVM LDA ELM	Channel pair C4-AF8 predicted 80% of mood labels (8 out of 10). Channel pair CPz-CP2 predicted 100% of cognitive labels (10 out of 10).	N/A

Arns, 2014	Linear and non-linear EEG analysis	90 subjects with MDD	rTMS and psychotherapy	DA	N/A	N/A
Erguzel, 2015	Pre-treatment Frontal QEEG	55 subjects with MDD -30 responders -25 non-responders	3 week, 20 sessions of rTMS as an adjunct to current medication. Stimulations were given to the lPFC	ANN Feed-forward neural network trained by backpropagation algorithm	89.09% (k=6) 85.45% (k=8) 87.27% (k=10)	k=6) TP 28, FN 2, TN 21, FP 4 (k=8) TP 26, FN 4, TN 21, FP 4 (k=10) TP 28, FN 2, TN 20, FP 5
Khodayari-Rostamabad, 2010a	Quantitative pre-clozapine resting EEG data	37 subjects with Schizophrenia (14 from independent sample)	Clozapine	RFS KPLSR LOOCV (Validation)	87.12% (for 52% response) 85.7% (for 30% response) 85.7% (>25% response)	No pre-treatment PANSS
Khodayari-Rostamabad, 2010b	Pre-treatment resting or spontaneous EEG data - 6 EEG files per patient (3 eyes open and 3 eyes closed)	22 subjects with MDD	6-week course of sertraline	KPLSR with GK * Feature selection with modified KL distance measure	86.60%	Sensitivity: 87.50% Specificity: 85.70%
Khodayari-Rostamabad, 2013	Pre-treatment EEG	22 subjects with MDD	6-week course of an SSRI (primarily Sertraline)	MDA * Feature selection using FDR	87.90%	Sensitivity: 94.86% Specificity: 80.93%
Mumtaz, 2017	Time frequency decomposition of EEG data	34 subjects with MDD	6-week course of an SSRI (escitalopram, fluvoxamine, sertraline or fluoxetine)	Feature extraction used multi-resolution decomposition of EEG data with WT analysis. STFT and EMD used for comparison. * Feature selection using 1) RBFS and 2) mRMR method	Model 1 - Wavelets Rank Based: 87.5%, mRMR, 72.0% Model 2 - STFT Rank Based: 80%, mRMR, 70.8% Model 3 - EMD Rank Based: 72.5%, mRMR 64.1%	Model 1 - Sensitivity: 95% (Rank based), 70% (mRMR) Specificity: 80% (Rank based), 77.5% (mRMR) Model 2 - Sensitivity: 75% (Rank based), 60% (mRMR) Specificity: 90% (Rank Based) 75% (mRMR)

					Model 4 (Best Model): Combination (Wavelets + STFT + EMD) Rank Based: 91.6%, mRMR: 76.2%	Model 3 - Sensitivity: 77.5% (Rank based), 60% (mRMR) Specificity: 67.5% (Rank based), 70% (mRMR) Model 4 - Sensitivity: 90% (Rank based), 72.5% (mRMR) Specificity: 90% (Rank based), 77.5% (mRMR)
First author, year	**Data utilized**	**Sample size and diagnosis[1]**	**Intervention**	**Machine learning model**	**Accuracy**	**Other measures**
STUDIES USING NEUROIMAGING						
Costafreda, 2009	Pattern of brain activity to sad faces in fMRI	16 patients with MDD, medication-free	16 sessions of cognitive behavioral therapy with experienced therapists	SVM with a linear kernel PCA to reduce dimensionality prior to analysis	NA	Sensitivity Lower/High sad intensity: 71% Mid-intensity: 57% Specificity Lower/High sad intensity: 86% Mid-intensity: 43%
Fleck, 2017	fMRI and proton magnetic resonance spectroscopy with a continuous performance task with emotional and neutral distractors	20 BD type I patients in with first-episode mania	8-week open-label lithium	LITHIA* LSVR RBFSVR SGD LAR LASSO RR	100% (classification) 89.8% (symptom reduction)	NA
Gong, 2011	sMRI	61 drug-naive patients with MDD	Single antidepressant drug with a minimum dose of 150mg/day of imipramine equivalents	SVM	GM: 69.57% WM: 65.22% Both: 69.57%	Sensitivity GM: 69.57% WM: 73.91% Both: 69.57% Specificity GM: 69.57% WM: 56.52% Both: 69.57%
Hahn, 2015	Whole brain fMRI during differential fear-conditioning task (acquisition and extinction	49 patients with PD with agoraphobia	12-sessions CBT twice a week	GPC	Acquisition: 73% Extinction: 74% Combined: 82%	Sensitivity Acquisition: 80% Extinction: 64%

	phases)					Combined: 92% Specificity Acquisition: 67% Extinction: 83% Combined: 72%
Liu, 2012	sMRI	36 MDD patients - 18 TRD - 17 TSD	Imipramine minimum dose of 150 mg/day or equivalent	FS: Searchlight+PCA, RFE, LLE Model: Linear SVM, C-Means	Gray Matter: 77.1-82.9% White Matter: 65.7-82.9%	NA
Mansson, 2015	BOLD-fMRI responses to self-referential criticism	26 patients with SAD	Internet delivered CBT and ABM	SVM	39-91.6%	AUC: 0.29-0.91 Sensitivity: 41.7-83.3% Specificity: 36.4-100%
Qin, 2015	Whole-brain rs-fcMRI	16 patients with MDD successfully treated	SNRI and SSRI antidepressant treatment	SVR t-test and PCA for feature selection	NA	MSE of 116.12 and correlation coefficient of -0.19 between real and predicted HDRS scores after treatment
Redlich, 2016	Whole-brain sMRI	23 patients with acute MDD	Brief-pulse ECT 3 time a week, 9 to 12 sessions, plus antidepressant treatment	SVM-L GPC	SVM: 78.3% GPC: 73.9%	Sensitivity SVM: 100% GPC: 100% Specificity: SVM: 50% GPC: 40%
Sikora, 2016	rs-fcMRI	29 patients with MDD	RCT of two weeklong, identical placebos described as having either “active” or “inactive”	ICA (unsupervised) RVR	NA	Resting-state functional connectivity of the SN was significantly predictive of placebo responses: correlation = .41; p = .018; mean sum of squares = 14.36; p = .019
Sundermann, 2017	fMRI based on an interoception task	59 patients with PD with agoraphobia	6 week randomized control trial consisting of manualized exposure-based CBT encompassing 12 x 100 min treatment sessions (two subgroups either with or without therapist-guided exposure)	Feature selection: t-test and SVM-RFE Model: SVM	38.0–54.2%	Sensitivity 30.0-50.0% Specificity 37.9-58.6%
Whitfield-Gabrieli, 2015	rs-fMRI and diffusion-weighted magnetic resonance imaging	38 patients with SAD	12 weekly sessions with CBT according to a standardized protocol-based group treatment	LR	81%	Sensitivity: 84% Specificity: 78%
First author, year	**Data utilized**	**Sample size and**	**Intervention**	**Machine learning**	**Accuracy**	**Other measures**

		diagnosis[1]		model		
STUDIES USING MULTIMODAL DATA						
Ball, 2014	Self-report clinical measures, fMRI emotion regulation task	48 subjects - 25 GAD - 23 PD	Open-label weekly individual CBT	RF DT	Clinical and demographic: 69% fMRI only: 79% Both: 73%	Sensitivity: Clinical and demographic: 79% fMRI only: 86% Both: 83% Specificity: Clinical and demographic: 53% fMRI only: 68% Both: 58%
Gowin, 2015	Clinical and neurocognitive assessment; fMRI during reward processing	68 patients with primary dependence on methamphetamine	28-day program including 12-step models, daily education and exercise, and Narcotics Anonymous meetings	RF DT	Clinical: 74% fMRI: 72% Both: 75%	Sensitivity Clinical: 67% fMRI: 61% Both: 75% Specificity Clinical: 77% fMRI: 77% Both: 81% AUC Clinical: 0.74 fMRI: 0.73 Both: 0.71
Hinton, 2017	Clinical data; metabolites (amino acids and amino acid derivatives) associated with treatment response to acamprosate	120 subjects with AUD	Acamprosate three times a day in a standard dose	LASSO	Training: 71.3% Test: 61.5%	AUC: 0.801 / 0.647 (Training/Test) Sensitivity: 94.9% / 83.3% Specificity: 36.6% / 31.0%
Hoogendoorn, 2016	Clinical and socio-demographic data; attributes extracted from e-mails	69 patients with SAD	Internet-based guided self-help intervention; therapists assistance and support via e-mail.	LR DT RF *Feature selection with Pearson correlation	NA	AUC LR: 0.71-0.83 / 0.55-0.75 CART: 0.62-0.70 / 0.45-0.64 RF: 0.72-0.81 / 0.48-0.78
Kim, 2015	Clinical and socio-demographic data; neuropsychological tests; genetic, environmental and neuroimaging measures	83 subjects with ADHD	8-week, open-label trial of MPH	SVM with 2nd order polynomial kernel DT RF LRR	(range) SVM: 64.1-84.6% DT: 61.5-69.2% RF: 61.5-73.1% LRR: 65.4-76.9%	AUC SVM: 0.55-0.84 DT: 0.51-0.61 RF: 0.58-0.79 LRR: 0.61-0.73
Lin, 2008	Five clinical variables and five pharmacogenetic variables	93 in-patients with schizophrenia	Clozapine for at least three months	ANN LR	Overall: ANN: 83.3%	AUC (Overall / Genetic only / Clinical only)

					LR: 70.8%	ANN: 0.821 / 0.805 / 0.647 LR: 0.579 / 0.516 / 0.604
Luo, 2014	Clinical, sociodemographic and behavioural data; PET scan from 5 regions of the striatum	25 subjects with cocaine dependence	Contingency management for 12 weeks, 3 clinic visits per week	LR SVM with radial basis function kernels	PET*4: 47-82% PET + Clinical: 74-77% PET + Behavioural: 68-96%	PET: 0.45-.87 PET + Clinical: 0.79-0.83 PET + Behavioural: 0.73-0.98
Patel, 2015	Clinical and sociodemographic data; rs-fMRI; sMRI	19 subjects with late-life depression	12-week open-label trial with duloxetine, venlafaxine, nimodipine, or escitalopram	L1-LR ADTree, Alternating SVM-L SVM-RBF Feature selection with Kendall tau correlation coefficient	Best model ADTree: 89.47%	Sensitivity ADTree: 88.89% Specificity ADTree: 90.00%
Rethorst, 2017	Clinical data and serum biomarkers (cytokines and neurotrophins)	122 patients with MDD	Patients were randomized to two exercise dose groups	LASSO RF	NA	AUC (average from both models) Remission: 0.785 Nonresponse: 0.710

Abbreviations:

AA, Arachidonic Acid; AAP, Atypical Antipsychotics; ABM, Attention Bias Modification; 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; BN, Bayesian Networks; BOLD, Blood-Oxygen Level-Dependent;* BSP, Brief Supportive Psychotherapy; BZD, Benzodiazepines; CBT, Cognitive Behavioural Therapy; DA, Discriminant Analysis; DF, Deterministic Forest; DHA, Docosahexaenoic Acid; 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; FF-BP ANN, Feed-forward Back-propagation Artificial Neural Network; *FDR, Fisher Discriminant Ratio; FS, Feature Selection; fMRI, Functional Magnetic Resonance Imaging; GAD, Generalized Anxiety Disorder; GB M, Gradient Boosting Machine; GK, Gaussian Kernel; GM, Gray Matter; GPC, Gaussian Process Classification; GPR, Gaussian Process Regression; ICA, Independent Component Analysis;* IPT-PS, Interpersonal Psychotherapy for Depression with Panic and Anxiety Symptoms; 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; 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; MPH, Methylphenidate; MRMR, Minimum redundancy and maximum relevance; NPV, Negative Predictive Value; OCD, Obsessive Compulsive Disorder; PCA, Principal Component Analysis; PD, Panic disorder; PHDD, Percentage of Heavy Drinking Days; PPV, Predictive Positive Value; PUFAs, Polyunsaturated Fatty Acids; SAD, Social Anxiety Disorder; SCG, Subcallosal Cingulate Gyrus; 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; RBFS, Rank-Based Feature Selection; RBFSVR, Radial Basis Support Vector Regression; 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;* 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. Akinci et al, 2013; Arribas et al, 2010; Redlich et al, 2014; and Valenza et al, 2013 didn't specify diagnostic criteria.
[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

Table 2: Quality Assessment of the included studies*

Paper	Representativeness of the sample	Control confounding	Assessment of the outcome	Machine learning algorithm	Performance metrics	Class imbalance	Test unseen	Feature selection + Hyper (2)	Missing data	Final Score
Al-Kaysi, 2016	0	0	0	1	0	0	0	0	0	1
Al-Kaysi, 2017	0	1	0	1	1	0	0	0	0	3
Amminger, 2015	1	1	1	1	1	0	0	0	1	6
Andreescu, 2008	1	1	0	0	0	0	0	1	1	4
Arns, 2014	1	1	0	1	1	0	0	0	1	5
Ball, 2014	0	1	0	1	1	1	0	1	1	6
Blankers, 2013	1	1	0	1	1	0	0	2	1	7
Block, 2010	1	1	0	1	1	0	1	0	1	6
Chekroud, 2016	1	0	1	1	1	0	1	1	1	7
Compton, 2005	1	1	0	1	1	1	1	1	1	8
Connor, 2007	0	0	0	1	1	0	1	0	1	4
Costafreda, 2009	1	1	0	1	1	1	1	0	1	7
Dodd, 2014	1	1	1	1	1	1	1	1	1	9
Doyle, 2014	1	1	0	1	1	1	0	1	1	7
Erguzel, 2015	1	1	0	1	1	1	0	0	0	5
Etkin, 2015	1	1	1	1	1	0	1	0	1	7
Fleck, 2017	0	1	0	1	1	0	0	0	0	3
Franchini, 2001	1	1	0	1	1	1	0	0	1	6
Gong, 2011	0	1	0	1	1	1	0	1	0	5
Gowin, 2016	1	1	0	1	1	0	0	0	0	4

Gueorguieva, 2015	1	1	0	0	1	0	1	1	1	6
Guilloux, 2015	1	1	1	1	1	0	1	2	0	8
Gupta, 2013	1	1	1	1	1	0	0	0	1	6
Han, 2015	1	1	0	1	1	1	0	0	0	5
Hannover, 2002	1	0	0	1	1	0	0	1	1	5
Hinton, 2017	1	1	0	1	1	1	1	1	0	7
Hoogendoorn, 2016	1	1	0	1	1	0	0	1	0	5
Hou, 2015	1	1	0	1	1	0	0	0	1	5
Iniesta, 2016	1	1	0	1	1	1	1	2	0	8
Johnston, 2015	0	1	0	1	1	1	0	1	1	6
Khodayari-Rostamabad, 2010	1	1	0	1	1	1	1	2	0	8
Khodayari-Rostamabad, 2010b	0	0	0	1	0	0	0	0	0	1
Khodayari-Rostamabad, 2013	1	1	1	1	1	1	1	2	1	10
Kim, 2015	1	1	0	1	1	0	0	1	0	5
Koutsouleris, 2016	1	1	1	1	1	1	1	2	1	10
Lin, 2008	1	1	1	1	1	0	1	1	0	7
Liu, 2012	0	1	0	1	1	0	0	1	0	4
Luo, 2015	1	1	0	0	1	1	1	1	0	6
Mansson, 2015	0	1	0	1	1	1	0	0	0	4
McInerney, 2017	1	1	0	0	1	1	0	0	0	4

Muller, 2008	1	0	0	1	0	0	0	1	0	3
Mumtaz, 2017	0	1	0	1	1	1	0	2	0	6
Nelson, 2012	1	1	0	1	1	0	1	1	1	7
Patel, 2015	0	0	1	1	1	0	1	2	1	7
Perlis, 2013	1	1	0	1	1	1	1	1	1	8
Pohjolainen, 2014	0	0	0	1	1	0	0	1	0	3
Politi, 2005	1	1	0	1	0	0	0	0	1	4
Qin, 2015	0	1	0	1	1	0	0	1	0	4
Quitkin, 1987	1	0	0	1	0	0	0	0	0	2
Redlich, 2016	1	1	1	1	1	0	0	0	0	5
Rethorst, 2017	1	1	0	1	1	0	0	2	1	7
Riedel, 2011	1	1	1	1	1	1	0	1	1	8
Ruberg, 2011	1	1	0	1	1	1	1	1	1	8
Salomoni, 2009	0	0	1	1	1	0	0	1	0	4
Serretti, 2007	1	1	1	1	1	1	1	1	1	9
Serretti, 2007b	1	1	1	1	1	0	0	0	1	6
Sikora, 2016	0	0	0	1	1	0	0	1	0	3
Stiles-Shields, 2016	1	1	1	1	1	0	1	0	0	6
Sundermann, 2017	0	1	1	1	1	1	0	2	0	7
Whitfield-Gabrieli, 2015	0	1	0	1	1	0	0	1	0	4
Winterer, 1998	0	1	0	1	0	0	1	1	1	5
Wong, 2017	1	1	0	1	1	1	1	1	0	7

* See Supplementary material for more information about the quality assessment instrument

1. Search Filter

("artificial intelligence" OR "machine learning" OR "learning machine" OR "machine intelligence" OR "computational learning" OR "computational intelligence" OR "computational psychiatry" OR "big data" OR "pattern recognition" OR "pattern classification" OR "predictive analysis" OR "predictive model" OR "classification model" OR "supervised learning" OR "semi supervised learning" OR "data science" OR "knowledge representation" OR "knowledge representations" OR "gaussian process" OR "regularized logistic" OR "linear discriminant analysis" OR "LDA" OR "random forest" OR "random forests" OR "naive bayes" OR "least absolute selection shrinkage operator" OR "elastic net" OR "LASSO" OR "support vector" OR "support vectors" OR "SVM" OR "RVM" OR "relevance vector machine" OR "neural network" OR "neural networks" OR "decision tree" OR "decision trees" OR "rule learner" OR "rule learners" OR "rule learning" OR "classification tree" OR "classification trees" OR "regression tree" OR "regression trees" OR "CART" OR "k-nn" OR "nearest neighbor" OR "nearest neighbour" OR "nearest neighbors" OR "nearest neighbours" OR "logistic regression" OR "multiple regression") AND (“mental disorder” OR "mental disorders" OR "psychiatric disorder" OR "psychiatric disorders" OR "anxiety disorder" OR "anxiety disorders" OR "obsessive compulsive" OR "panic disorder" OR "phobic disorder" OR "social phobia" OR "social anxiety" OR "generalized anxiety disorder" OR "GAD" OR "specific phobia" OR "post traumatic stress disorder" OR "PTSD" OR "gambling" OR "anorexia" OR "bulimia" OR "binge eating" OR "mood disorder" OR "mood disorders" OR "affective disorder" OR "affective disorders" OR "major depressive disorder" OR "depressive disorder" OR "bipolar disorder" OR "schizoaffective" OR "schizophrenia" OR "psychotic disorder" OR "psychotic disorders" OR "attention deficit" OR "suicide" OR "substance use disorder" OR "alcohol" OR "heroin" OR "cocaine" OR "personality disorder") AND (treatment OR intervention OR therapeutic OR therapy OR "side effects" OR "undesirable effects" OR "injurious effects" OR "adverse effects")

2. Quality assessment instrument development

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

We considered the methodological features comprising sample representativeness, confounding variables, and outcome assessments, as the most clinically relevant aspects among machine learning based healthcare research. Relevant considerations of each methodological feature are discussed in further detail within the table. 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 results were optimized using hyper-parameter optimization and feature selection procedures, 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 accuracy

was tested in unseen data.

3. Quality assessment instrument domains

Methodological Feature	Considerations
1. Representativeness of the sample	Was the study truly representative of the target population heterogeneity? If not, was this related to the selected 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. Machine Learning Approach	Was the machine learning algorithm used to analyze 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 if they were inputted or removed?
8. Performance/Accuracy	Were the following performance metrics included: A. Accuracy B. Sensitivity C. Specificity D. AUC E. PPV/NPV
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?

6.3. Artigo 3

Título: **"Prediction of depression incidence, remission, and persistence in a large occupational cohort using machine learning techniques: an analysis of the ELSA-Brasil study"**

Submetido à *Neuropsychopharmacology* em agosto de 2019, atualmente em revisão.

Fator de impacto da revista: 7.160

Gmail

Diego Librenza Garcia <librenzagarcia@gmail.com>

NPP-19-0943 Receipt of New Paper by Neuropsychopharmacology

journal@acnp.org <journal@acnp.org> 6 August 2019 at 09:32
Reply-To: journal@acnp.org
To: librenzagarcia@gmail.com

Dear Dr Librenza Garcia,

Please note that you are listed as a co-author on the manuscript "Prediction of depression incidence, remission, and persistence in a large occupational cohort using machine learning techniques: an analysis of the ELSA-Brasil study" (reference number: NPP-19-0943), which was recently submitted to Neuropsychopharmacology.

Please take this opportunity to login to the submission portal now and review your Funding and Disclosure details as you described in the manuscript. If you have any changes to this, please alert the corresponding author immediately and ensure these are in line with the journal's policies. It is the authors' responsibility to disclose relevant interests in the work.

The corresponding author is solely responsible for communicating with the journal and managing communication between co-authors. Please contact the corresponding author directly with any queries you may have related to this manuscript.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals. Please check your account regularly and ensure that we have your current contact information.

In addition, NPG encourages all authors and reviewers to associate an Open Researcher and Contributor Identifier (ORCID) to their account. ORCID is a community-based initiative that provides an open, non-proprietary and transparent registry of unique identifiers to help disambiguate research contributions.

Access your account

Many thanks,
NPG Applications Helpdesk
Springer Nature Limited

NPP - This email has been sent through the Springer Nature Tracking System NY-610A-NPG&MTS

Prediction of depression incidence, remission, and persistence in a large occupational cohort using machine learning techniques: an analysis of the ELSA-Brasil study

Authors: Diego Librenza-Garcia[1,2]; Ives Cavalcante Passos[1]; Jacson Feiten[1]; Paulo A Lotufo[3,4]; Alessandra C Goulart[3,4]; Itamar de Souza Santos[3,4]; Maria Carmen Viana[5]; Isabela M Benseñor[3,4]; Andre Russowsky Brunoni[3,4,6]

(1) Laboratory of Molecular Psychiatry, Hospital de Clínicas de Porto Alegre; Programa de Pós-Graduação em Psiquiatria e Ciências do Comportamento, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil;
(2) Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada.
(3) Department of Internal Medicine, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.
(4) Hospital Universitário, Universidade de São Paulo, São Paulo, Brazil.
(5) Center of Psychiatric Epidemiology (CEPEP), Department of Social Medicine, Postgraduate Program in Public Health, Federal University of Espírito Santo, Vitória, Brazil
(6) Laboratory of Neurosciences (LIM-27), Department and Institute of Psychiatry, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil.

Running Title: Prediction of incidence, remission, and persistence of depression.

Corresponding Author: André Russowsky Brunoni, MD, PhD, Av. Prof Lineu Prestes 2565, Hospital Universitário (HU-USP), CEP 05508-000, Butantã, São Paulo, Brazil, e-mail: brunoni@usp.br

Email for authors:

Diego Librenza-Garcia: librenzagarcia@gmail.com

Ives Cavalcante Passos: ivescp1@gmail.com

Jacson Feiten: jacsonfeiten95@gmail.com

Paulo Lotufo: palotufo@usp.br

Alessandra C Goulart: agoulart@hu.usp.br

Itamar de Souza Santos: itamarss@usp.br

Maria Carmen Viana: mcviana6@gmail.com

Isabela M Benseñor: isabensenor@gmail.com

André Russowsky Brunoni: brunoni@usp.br

Word count: words

Abstract word count: 271 words

Abstract

Depression is highly prevalent and marked by a chronic and recurrent course. Despite being a major cause of disability worldwide, little is known regarding determinants of its heterogeneous course. Machine learning techniques present an opportunity to develop tools to predict diagnosis and prognosis at an individual level. In the present study, we examined baseline (2008-2010) and follow-up (2012-2014) data of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil), a large occupational cohort study. We implemented an elastic net regularization analysis using socioeconomic and clinical factors as predictors to distinguish: (1) depressed from non-depressed participants, (2) participants with incident depression from those who did not develop depression, and (3) participants with persistent depression from those without depression. At wave 1 and 2, we assessed 15,105 and 13,922 participants, respectively. The elastic net regularization model distinguished outcome levels in the test dataset with an area under the curve of 0.90 (95% CI 0.85 - 0.95), 0.89 (95% CI 0.85 - 0.94), 0.90 (95% CI 0.86 - 0.95) for analyses 1, 2, and 3, respectively. We conclude that diagnosis and prognosis related to depression can be predicted at an individual subject level by integrating demographic and clinical variables, such as psychiatric comorbidities. Future studies should assess longer follow-up periods and combine biological predictors, such as genetics and blood biomarkers, to build more accurate tools to predict depression course.

Keywords: major depressive disorder; prognosis; machine learning; incident depression.

1. Introduction

Mood disorders account for almost 50% of the burden of mental disorders [1]. Among them, major depression has a chronic, recurrent course and is highly prevalent. In fact, its chronicity is especially difficult to tackle, as research is usually focused on the management of acute depressive episodes [2]. In addition, depression has a heterogeneous development, varying from a single or a few episodes to an intermittent course that can persist over the lifespan [3].

However, relatively little is known on the sociodemographic and clinical predictors associated with depression recurrence and incidence, as most available cohort data present limitations. For instance, several cohort studies investigated specific subgroup of patients (e.g., perinatal depression, geriatric depression, and depression in children and adolescents), enrolled only those already depressed at baseline, performed a short-term follow-up, or presented high attrition rates [3–6]. An additional issue is that most studies have been conducted in developed countries [3]. It is reasonable to assume that the course of depression is different in low- and middle-income countries that present substantial economic disparity and low social support for the poorest people.

Importantly, standard investigation has focused on traditional statistical approaches that explore a linear relationship between variables at group-level data [7]. In this context, machine learning approaches can be advantageous and have increasingly been used in prognostic psychiatry, as they can assume a complex relationship between variables, including non-linear patterns, and are focused at an individual patient level [8]. For example, there is supporting evidence for the influence of socioeconomic inequality in the association between depression and gender [9].

Considering these issues, we evaluated, using a machine learning approach, predictors of depression incidence, persistence, and remission in a large Brazilian occupational cohort. This study has the potential of influencing mental health policies. Additionally, our findings could be employed in other low- and middle-income countries that present populations with similar characteristics.

2. Methods

2.1. Study design and participants

ELSA-Brasil is a prospective, occupational cohort study of 15,105 civil servants from six public institutions in major Brazilian cities (São Paulo, Rio de Janeiro, Salvador, Porto Alegre, Belo Horizonte and Vitória) [10]. All active or retired employees of these institutions aged 35–74 years were eligible for the study. Exclusion criteria were current or recent pregnancy (4 months prior to the first interview), intention to quit working at the institution in the near future, severe cognitive or communication impairments, and, if retired, residence outside of a study center's corresponding metropolitan area. All local ethics committees approved the study and all participants provided written, informed consent prior to assessment.

The first wave (n=15,105 participants) of ELSA took place from August 2008 to December 2010 and the second wave (n=13,922 participants) took place from September 2012 to December 2014.

2.2. Predictor variables

As predictors, we selected variables that are easily accessible to clinicians and that can be collected in a single clinic visit. In consequence, the models created can be used in large populations without significative increase in the cost of assistance.

The following baseline variables were investigated as predictors:

a) for sociodemographic variables, information was collected regarding sex, age, educational level (presence or absence of a university degree), self-reported race (white vs. non-white), marital status (married vs. other), and familial monthly income.

b) regarding clinical variables, we assessed obesity (defined as a body mass index > 30 kg/m^2 and obtained by measured weight and height) and smoking status (never a smoker vs. past or present smoker). To evaluate general health status, participants were asked to judge their health according to a Likert scale. The answers were categorized into very good/good health status vs. moderate/poor/very poor health. Finally, we used dietary information to identify those who presented a heavy alcohol consumption [11], defined as more than 210 (men) or 140 (women) grams of alcohol consumed per week [12].

c) regarding mental disorders, we used the Portuguese version of the Clinical Interview Schedule-Revised (CIS-R) [13], which is a structured interview for measurement and diagnosis of non-psychotic psychiatric morbidity in the community [14]. The questionnaire includes 14 sections covering common psychiatric symptoms and assessing the following ICD-10 diagnosis: general anxiety disorder (GAD, F41.1), panic disorder (PD, F41.0), social anxiety disorder (SAD, F40.1), and obsessive-compulsive disorder (F42).

d) for psychotropic use, all participants were asked regarding use of prescription and nonprescription medicines, and continuous and non-continuous use of medication taken in the past two weeks. All participants were instructed to bring to the study clinic all medications and prescription forms for examination. We assessed whether participants were using antidepressants and/or benzodiazepines. A complete review on psychotropic use in the ELSA-Brasil study can be found elsewhere [15].

e) finally, we assessed the presence or absence of at least one of the following negative life events in the past 12 months: being assaulted or robbed, being hospitalized, bereavement/mourning of a relative, severe financial problems, or ending an intimate relationship.

2.3. Outcome variables

At follow-up (wave 2), a shortened version of the CIS-R was applied to diagnose depression. Therefore, we could define the following clinical courses: no depression (absence of depression at both waves), incident depression (depression only at wave 2), remitted depression (depression only at wave 1), and persistent depression (depression at both waves).

2.4. Data Analysis

Descriptive analyses were reported as means (with standard deviations) or absolute and relative frequencies. We divided participants into four groups based on the outcomes. We used chi-squared ($\chi 2$) or Student t-tests to analyze demographic and clinical variables among these groups.

The machine learning analysis was performed with R software (Version R 3.3.1) and R Studio (Version 0.99.902) using the R package caret (Version 6. 0-73) [16]. Machine learning approaches are superior to traditional multiple regression

analyses because: 1) coefficients are unstable when high correlations exist among predictors, which leads to low replication of predictions in independent samples [17]; 2) traditional regression assumes additivity, whereas the predictors considered here might have non-additive effects.

2.5. Machine learning analysis

The elastic net is a machine learning method that uses regularization with an embedded feature selection procedure. Through a cost function composed of both L_1 (Lasso regression) and L_2 (Ridge regression) weight magnitude penalties, the method is able to remove predictors with low impact to the outcome while regularizing for improved generalization. The coefficients of the non relevant features are shrunk towards zero, eliminating correlated variables, simplifying the model, and reducing overfitting. As our dataset is composed of several attributes, identifying the most important ones enables a wider applicability and more practical use of our predictive models. We performed bivariate elastic net regularization to explore the association of the predictive variables and the outcome. The following outcomes at wave 2 were considered: 1) no depression vs. any type of depressive course (incident or persistent or remitted); 2) no depression vs. incident depression; and 3) no depression vs. persistent depression. We performed the missing data imputation by using median for numeric variables and mode for categorical variables, using the training dataset [18].

Individual-level predicted probabilities based on the elastic net algorithm were created, as well as the receiver operating characteristic (ROC) curves, and the area under the curve (AUC) was calculated to evaluate the predictive performance. Additionally, we calculated sensitivity, specificity, balanced accuracy, positive predictive value (PPV), and negative predictive value (NPV). We used a cut-off of 0.5 as the boundary for class decision, i.e., the algorithm will classify probabilities above 50% as belonging to the positive outcome level and below to the negative outcome level. Finally, we plotted how PPV and NPV changes vis-à-vis different cut-offs for class boundary decision.

Cross-validation

For each analysis, we randomly split our baseline data into training (75% of the whole sample) and test datasets (25%). We deployed a standard machine

learning protocol with 10-fold cross-validation, feature selection, hyperparameter tuning, and class imbalance correction in the training dataset (Figure 1). We repeated 10-fold cross-validation ten times to improve tuning.

Class imbalance

Class imbalance introduces a bias towards classifying all the data as the majority class, which usually leads to poor detection of the infrequent class. The class imbalance problem was addressed through a resampling step, which entailed under-sampling the majority class in each analysis followed by algorithm training. This process was repeated in 1000 iterations to allow us to use all instances in the training set. The algorithm predicted probabilities were averaged over the resampling iterations. When dealing with imbalanced datasets, AUC and balanced accuracies are the best performance metrics to represent the results.

3. Results

Out of the 15,105 participants included at wave 1, 1180 (7.8%) did not complete the assessment at wave 2, the main reasons being death and moving outside of the metropolitan area of the study after retiring. We found that 499 (3.58%) participants presented with a new depressive episode, 426 (3.06%) remitted, 160 (1.15%) persisted in a depressive episode, and 12,837 (92.21%) presented no current depression at the follow-up. Descriptive analyses of demographic and clinical variables are described in Table S1, and missing data frequency and distribution for each variable is presented in Figure S1 of the supplementary material.

Figure 2 shows the ROC curves and AUC values for the predictive models regarding outcomes 1-3, and the selected variables with their relative relevance weights to each model. Figure S2 of the supplementary material shows PPVs and NPVs values for different cut-offs of class boundaries. Table 1 shows model performance for each analysis when the cut-off is chosen as 0.5.

3.1. Classifying depressed and non depressed patients

Considering both baseline and follow-up, 1,085 participants presented with a history of depression, while 12,387 participants have not experienced any depressive episode. The elastic net model had an AUC of 0.90 (0.85-0.95) with a balanced accuracy of 81%. The model retained all variables except past or present history of smoking. In the five top features selected, there were four comorbidities (SAD, OCD, GAD and PD) and the self-reported health evaluation.

3.2. Prediction of incident depression

There were a total of 499 participants with a new depressive episode at follow-up. The model was trained to differentiate these incidents cases from the non depressed patients at wave 2. The model had an AUC of 0.89 (0.85-0.94) and a balanced accuracy of 79%. Among the five top variables, there were two comorbidities (OCD and GAD), two clinical features (use of antidepressants and use of benzodiazepines) and sex. Past or present history of smoking and having a university degree were discarded by the model.

3.3. Distinguishing persistent depression from non depressed patients

At wave 2, 160 patients that were depressed at wave 1 persisted in a depressive episode. The model was trained to differentiate persistent depressed participants from those without a depressive episode, and had an AUC of 0.90 (0.86-0.95), with a balanced accuracy of 82%. OCD and GAD were the most relevant features, with sex, self-report health and negative life events following. The only variable discarded by the model was the self-reported race.

3.4. Sensitivity analyses

Since the use of antidepressants may be a confounder, we repeated the analysis for outcomes 1 and 3 without this variable to check if the performance could be inflated by its inclusion. The results can be seen in Figure 3 and Table 2. The model to distinguish depressed from non depressed patients had a decrease in the AUC from 0.90 (0.85-0.95) to 0.79 (0.78-0.81), while the model to distinguish non depressed participants from those with persistent depression had an absolute increase from 0.90 (0.86-0.95) to 0.91 (0.89-0.94), although there was a significant overlap in the confidence intervals. The same two variables previously excluded (history of smoking for model 1 and ethnicity for model 3) were also excluded in the sensitivity analysis models.

4. Discussion

The present study investigated three predictive models for incidence, remission, and persistence of depression within the ELSA-Brasil cohort, using baseline variables from wave 1 (2008-2010) as predictors, and outcomes measured at wave 2 (2012-2014). The present study is the first to assess depression prognosis in a large sample using machine learning techniques. Particularly, we designed predictive models to distinguish a) participants with depression from those without depression; b) participants with incident depression from those without depression; and c) participants with persistent depression from those without depression. We obtained AUCs ranging from 0.89 to 0.90, and balanced accuracies ranging from 79 to 82%.

Predicting which individuals are at-risk to convert to depression can enable timely and personalized preventive strategies to take place, shifting our focus from only treating acute episodes to directly intervening in the course of the disorder. This may yield a substantial impact to ease the burden directly associated with depression, such as cognitive and functioning impairments [19], high risk for suicidal behavior [20,21], and decreased quality of life [22]. In addition, it could impact also in mortality and disability rates, as well as in the economic and family burden associated with the disorder [23]. For example, depression is a risk factor for clinical diseases such as diabetes [24], coronary heart disease [25,26], and autoimmune diseases[27], with patients being twice as likely to die prematurely when compared to subjects without depression [19]. Our findings show a potential application of machine learning in predicting incidence, persistence and remission of depressive episodes at an individual level. In addition, the models developed in this study are easy to implement, since all variables can be accessed at any moment by a clinician, without incurring in additional costs. Interventions focused on the course of the disorder can be designed to target the relevant factors selected in the predictive models, since from all included variables, only age, sex and ethnicity are not modifiable [27].

Comorbidities were between the most relevant predictive features in all models. This is in accordance to a previous study by our group, that showed large effects sizes for OCD and anxiety disorders to predict incident and persistent depression, using traditional statistical methods [28]. The present study differs from our previous one

as we employed more variables, tested more outcomes, and used a machine learning approach. Our findings are also in accordance with a recent meta-analysis of 66 prospective studies that showed that anxiety disorders predict depressive disorder, with effect sizes of 2.58 [1.81, 5.2] for GAD, 2.06 [1.71, 3.97] for SAD, and 5.60 [4.21, 6.01] for OCD [29]. Some authors also consider the presence of high anxiety traits as a phenotype with increased predisposition to stress-induced depression [30]. While SAD was the most important feature to differentiate depressed from non-depressed patients, it had an intermediate relevance for the other outcomes.

Regarding medication use, the use of benzodiazepines had an intermediate relevance for the three models. The use of antidepressants had an intermediate relevance for most models, except for the one predicting incident depression, in which it was the fourth more relevant feature. Of note, we decided to include use of antidepressants in all our models because patients may be using these for other reasons than being depressed, such as treatment of chronic pain, anxiety disorders and obsessive compulsive disorder. For example, there is evidence that subjects with GAD not treated with antidepressants have a higher risk to develop a depressive episode [31]. When performing a sensitivity analysis removing this variable for models 1 and 3, the first model had a mild decrease in performance, while the third remained at a similar value.

Among all sociodemographic features, the most relevant for all models was sex, being the third most relevant feature to predict incident and persistent depression, and the sixth more relevant to distinguish between depressed and non depressed patients. The role of sex in depression is well-known, with women being twice as likely to develop depression [32], although there is no conclusive evidence of its role in remission, recurrence or persistence [33]. The difference in incidence seems to be higher during adolescence, period which was not included in our population. Having a university degree was the sixth more important variable to predict persistent depression and to distinguish remitted from persistent course, although it was discarded by the model of incident depression. Other sociodemographic variables had an intermediate to small relevance in the models. Age, for example, was among the five less relevant features in all models. This could be explained by the age

range of our sample (35-74 years) and the fact the incidence is higher during adolescence and early adulthood [34].

Our study had some limitations. Since this is an occupational cohort, it is uncertain if the findings can be generalized to a community sample. Due to the nature of the sample, unemployment, that is known as a risk factor for both depression and a more pernicious course of depressive symptoms, could not be included as a predictor. Although the lack of a large set of features can be considered a limitation, since other variables could improve further the model performance, it can also be seen as an advantage, since a small set of features that are easy and fast to collect makes a more feasible tool, that can be used in large populations with small costs. In addition, we had a large sample size available, which makes the machine learning process more robust. Finally, an important limitation is the short follow-up period, which may have influenced the high rates of false positives found, and the fact that patients were only assessed in two points in time, while for a more reliable determination of depressive trajectories, more evaluations and for longer periods of time are required.

5. Conclusion

In the present study, we developed three predictive models of depressive course in an occupational cohort, using machine learning techniques. Using a small number of clinical and sociodemographic predictors, we showed that it is possible to distinguish non depressed participants from those with depression, including incident and persistent cases, with high model performance. In addition, we also showed that clinical variables seem to be, at least for this sample, more relevant than sociodemographic variables. Knowing beforehand which individuals will have a depressive episode, and within these, which will have a more chronic and debilitating course, could help improve how we assess patients in clinical settings, shifting our focus from treating acute episodes, to preventing them.

References

1. Kupfer DJ, Frank E, Phillips ML. Major depressive disorder: new clinical, neurobiological, and treatment perspectives. Lancet. 2012;379:1045–1055.

2. Andrews G. Reducing the Burden of Depression. Can J Psychiatry. 2008;53:420–427.

3. Musliner KL, Munk-Olsen T, Eaton WW, Zandi PP. Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. J Affect Disord. 2016;192:199–211.

4. Spijker J, de Graaf R, Bijl R V, Beekman ATF, Ormel J, Nolen WA. Determinants of persistence of major depressive episodes in the general population. Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). J Affect Disord. 2004;81:231–240.

5. Beard JR, Tracy M, Vlahov D, Galea S. Trajectory and Socioeconomic Predictors of Depression in a Prospective Study of Residents of New York City. Ann Epidemiol. 2008;18:235–243.

6. Skapinakis P, Weich S, Lewis G, Singleton N, Araya R. Socio-economic position and common mental disorders. Br J Psychiatry. 2006;189:109–117.

7. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15:233–234.

8. Dwyer DB, Falkai P, Koutsouleris N. Machine Learning Approaches for Clinical Psychology and Psychiatry. Annu Rev Clin Psychol. 2018;14:91–118.

9. Rai D, Zitko P, Jones K, Lynch J, Araya R. Country- and individual-level socioeconomic determinants of depression: multilevel cross-national comparison. Br J Psychiatry. 2013;202:195–203.

10. Aquino EML, Barreto SM, Bensenor IM, Carvalho MS, Chor D, Duncan BB, et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): Objectives and Design. Am J Epidemiol. 2012;175:315–324.

11. Chor D, Alves MG de M, Giatti L, Cade NV, Nunes MA, Molina M del CB, et al. Questionnaire development in ELSA-Brasil: challenges of a multidimensional instrument. Rev Saude Publica. 2013;47:27–36.

12. Piccinelli M, Tessari E, Bortolomasi M, Piasere O, Semenzin M, Garzotto N, et al. Efficacy of the alcohol use disorders identification test as a screening tool for hazardous alcohol intake and related disorders in primary care: a validity study. BMJ. 1997;314:420–420.

13. Nunes MA, Pinheiro AP, Bessel M, Brunoni AR, Kemp AH, Benseñor IM, et al. Common mental disorders and sociodemographic characteristics: baseline findings of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). Rev

Bras Psiquiatr. 2016;38:91–97.

14. Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. Psychol Med. 1992;22:465–486.

15. Brunoni AR, Nunes MA, Figueiredo R, Barreto SM, Fonseca M de JM da, Lotufo PA, et al. Patterns of benzodiazepine and antidepressant use among middle-aged adults. The Brazilian longitudinal study of adult health (ELSA-Brasil). J Affect Disord. 2013;151:71–77.

16. RC T. R: A language and environment for statistical computing.

17. Berk RA. Statistical Learning from a Regression Perspective. 2016.

18. Van Buuren S. Van Buuren S. Flexible imputation of missing data. Chapman and Hall/CRC, 2018e.

19. Lépine J. The increasing burden of depression. 2011;7:3–7.

20. Bostwick JM, Pankratz VS. Reviews and Overviews Affective Disorders and Suicide Risk: A Reexamination. Am J Psychiatry. 2000;157:1925–1932.

21. Ösby U, Brandt L, Correia N, Ekbom A, Sparén P. Excess Mortality in Bipolar and Unipolar Disorder in Sweden. Arch Gen Psychiatry. 2001;58:844.

22. Brenes GA. Anxiety, depression, and quality of life in primary care patients. Prim Care Companion J Clin Psychiatry. 2007;9:437–443.

23. Kessler RC. The Costs of Depression. Psychiatr Clin North Am. 2012;35:1–14.

24. Brown LC, Majumdar SR, Newman SC, Johnson JA. History of Depression Increases Risk of Type 2 Diabetes in Younger Adults. Diabetes Care. 2005;28:1063–1067.

25. O'Neil A, Fisher AJ, Kibbey KJ, Jacka FN, Kotowicz MA, Williams LJ, et al. Depression is a risk factor for incident coronary heart disease in women: An 18-year longitudinal study. J Affect Disord. 2016;196:117–124.

26. Gan Y, Gong Y, Tong X, Sun H, Cong Y, Dong X, et al. Depression and the risk of coronary heart disease : a meta-analysis of prospective cohort studies. 2020:1–11.

27. Andersson NW, Gustafsson LN, Okkels N, Taha F, Cole SW. Depression and the risk of autoimmune disease : a nationally representative , prospective longitudinal study. 2017:3559–3569.

28. Brunoni AR, Santos IS, Passos IC, Goulart AC, Koyanagi A, Carvalho AF, et al. Socio-demographic and psychiatric risk factors in depression trajectories: a cohort analysis in ELSA-Brasil. Unpublished. 2019. 2019.

29. Jacobson NC, Newman MG. Anxiety and depression as bidirectional risk

factors for one another: A meta-analysis of longitudinal studies. Psychol Bull. 2017;143:1155–1200.

30. Weger M, Sandi C. High anxiety trait: A vulnerable phenotype for stress-induced depression. Neurosci Biobehav Rev. 2018;87:27–37.

31. Goodwin RD, Gorman JM. Psychopharmacologic Treatment of Generalized Anxiety Disorder and the Risk of Major Depression. Am J Psychiatry. 2002;159:1935–1937.

32. Kuehner C. Gender differences in unipolar depression: an update of epidemiological findings and possible explanations. Acta Psychiatr Scand. 2003;108:163–174.

33. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. Psychol Bull. 2017;143:783–822.

34. Saluja G, Iachan R, Scheidt PC, Overpeck MD, Sun W, Giedd JN. Prevalence of and Risk Factors for Depressive Symptoms Among Young Adolescents. Arch Pediatr Adolesc Med. 2004;158:760.

Funding and disclosure

The ELSA-Brasil study was supported by the Brazilian Ministry of Health and CNPq (grants 01060010.00RS, 01060212.00BA, 01060300.00ES, 01060278.00MG, 01060115.00SP, 01060071.00RJ)

The authors declare that they have no competing interests.

Figure 1: Elastic net procedure for training and testing data.

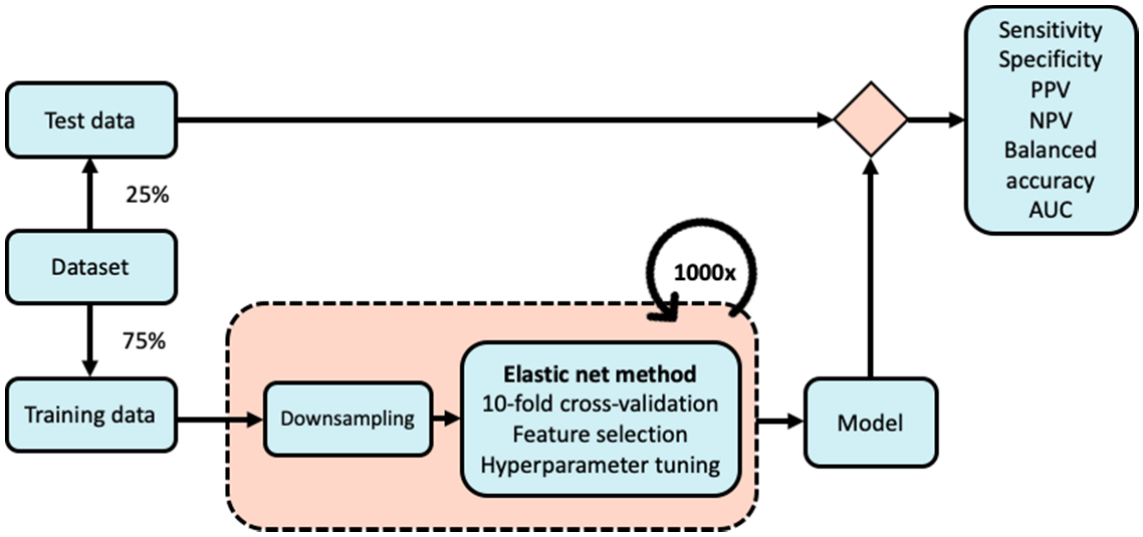
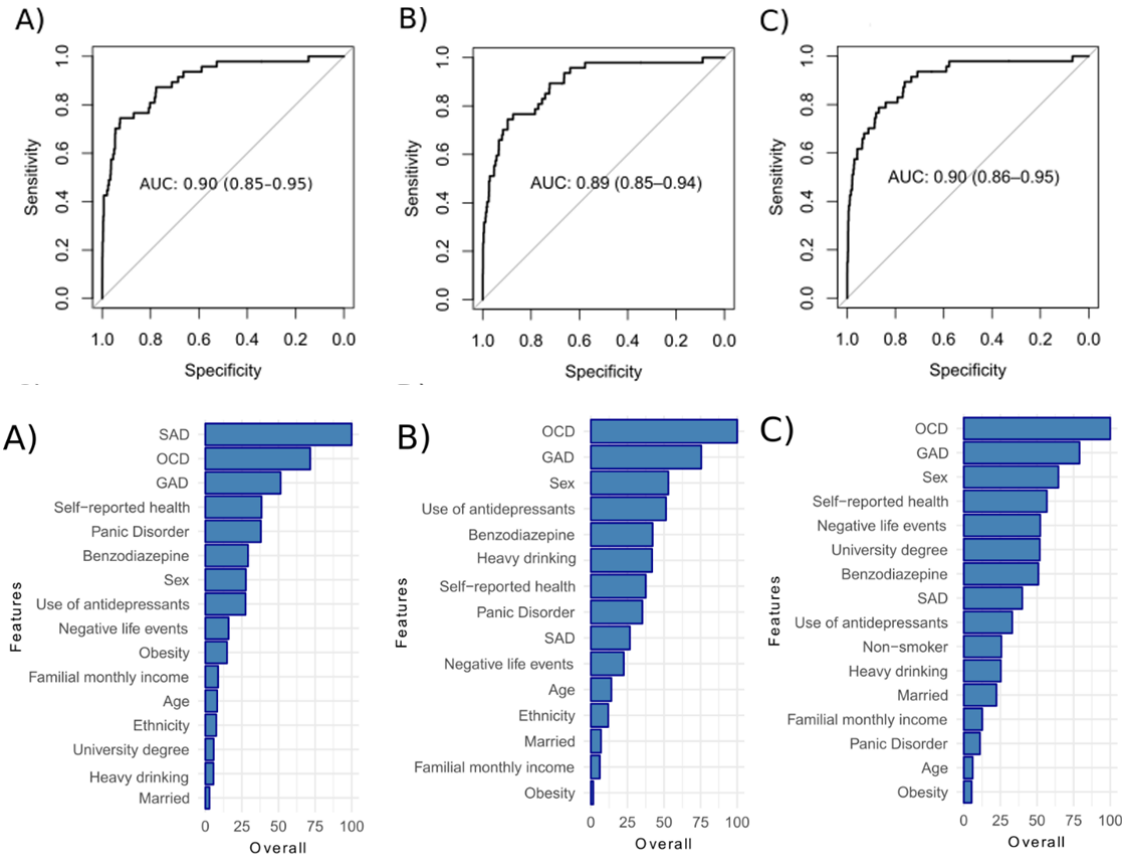
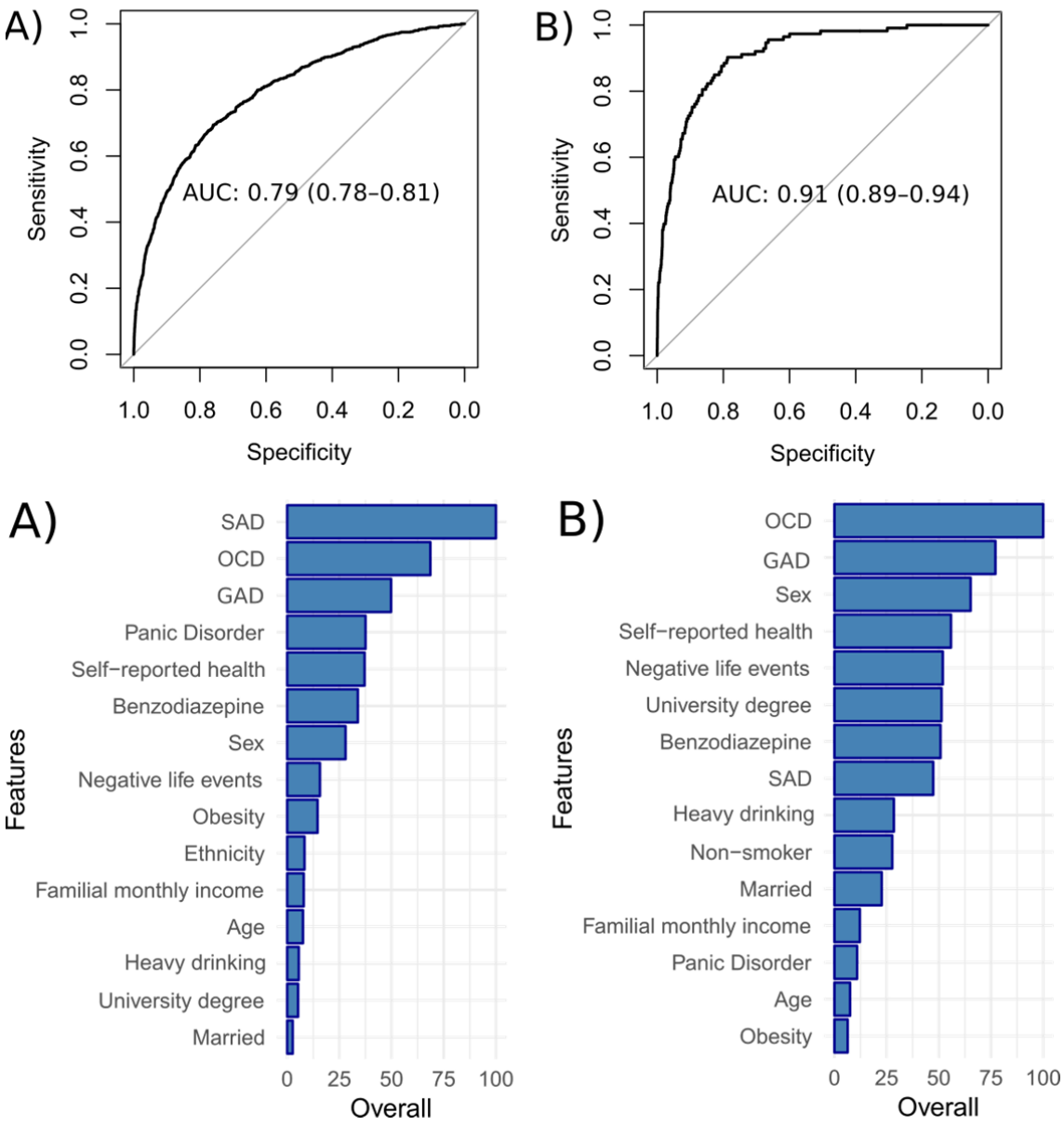


Figure 2: ROC curve and AUC value for the predictive models of depression courses and variables selected by the elastic net model with relative relevance weights.



Models differentiating (A) participants with depression from non-depressed participants; (B) participants with incident depression from participants who did not develop depression; (C) participants without depression from those with persistent depression.

Figure 3: ROC curves and selected variables with their relative relevance weights for sensitivity analysis for outcomes 1 and 3.



A) Depressed versus non depressed patients. B) Persistent depression versus non depressed patients.

Table 1: Performance metrics for the elastic net models to predict the three clinical outcomes.

Model	Sensitivity	Specificity	PPV	NPV	Balanced Accuracy	AUC
A	0.83	0.78	0.05	1.00	0.81	0.90 (0.85 - 0.95)
B	0.85	0.74	0.05	1.00	0.79	0.89 (0.85 - 0.94)
C	0.81	0.84	0.07	1.00	0.82	0.90 (0.86 - 0.95)

A) depression versus non depression; B) incident depression versus non depression; C) persistent depression versus non depression. AUC, area under the curve; NPV, negative predictive value; PPV, predictive positive value.

Table 2: Performance metrics for the sensitivity analysis for outcomes 1 and 3.

Sensitivity Analysis (SA)

Model	Sensitivity	Specificity	PPV	NPV	Balanced Accuracy	AUC
A-SA	0.67	0.78	0.21	0.96	0.72	0.79 (0.78 - 0.81)
C-SA	0.83	0.83	0.06	1.00	0.83	0.91 (0.89 - 0.94)

Comparison 1

	No-depression	Depression
No antidepressants	12143	921
Antidepressants	694	164

X-squared = 161.4 p-value < 2.2e-16

Comparison 3

	No persistent depression	Persistent depression
No antidepressants	12143	126
Antidepressants	694	34

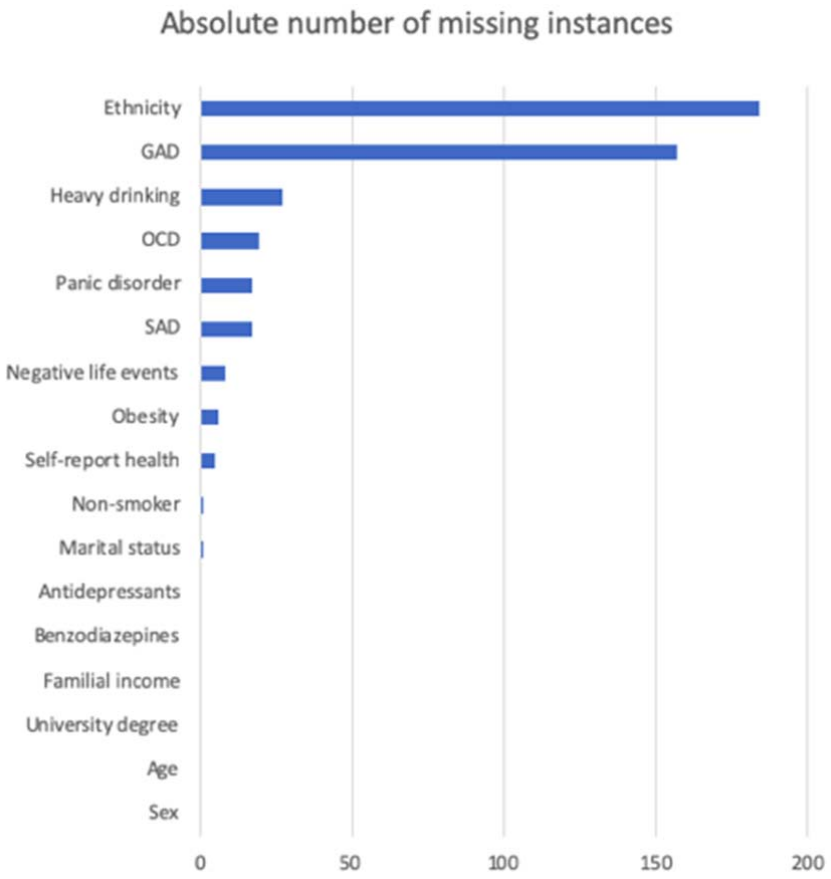
X-squared = 72.06 p-value < 2.2e-16

Supplementary Material

Table S1: Descriptive analyses of demographic and clinical variables

	All sample	Non depressed	Incident	Remitted	Persistent
	(*n=13922*)	(*n=12837*)	(*n=499*)	(*n=426*)	(*n=160*)
Age (mean±SD)	51.83 (±8.98)	51.94 (±9.02)	50.36 (±8.28)	50.61 (± 8.30)	50.86 (± 8.41)
Sex (female)	7597 (54.6)	6805 (53.0)	352 (70.5)	308 (72.3)	132 (82.5)
Ethnicity (Caucasian)	7220 (52.4)	6729 (53.0)	226 (46.2)	201 (47.4)	64 (40.0)
University degree (yes)	7443 (53.5)	6992 (44.3)	221 (44.3)	180 (42.3)	50 (31.2)
Familial Monthly income (mean±SD)	1748.00 (1437.9)	1782.21 (1456.22)	1443.60 (1143.58)	1297.85 (1110.44)	1154.24 (1074.91)
Marital status (Living with partner vs. Other)	9239 (66.4)	8617 (57.1)	303 (60.7)	237 (55.6)	82 (51.2)

Figure S1: Missing data distribution in absolute instances missed and percentage of missing data per variable.



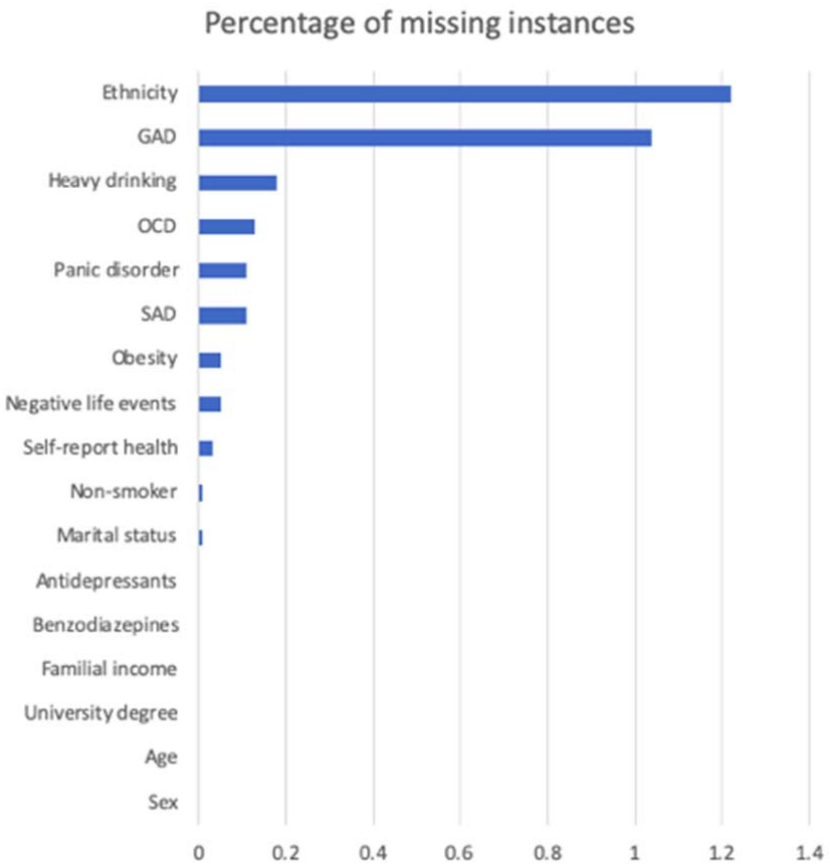
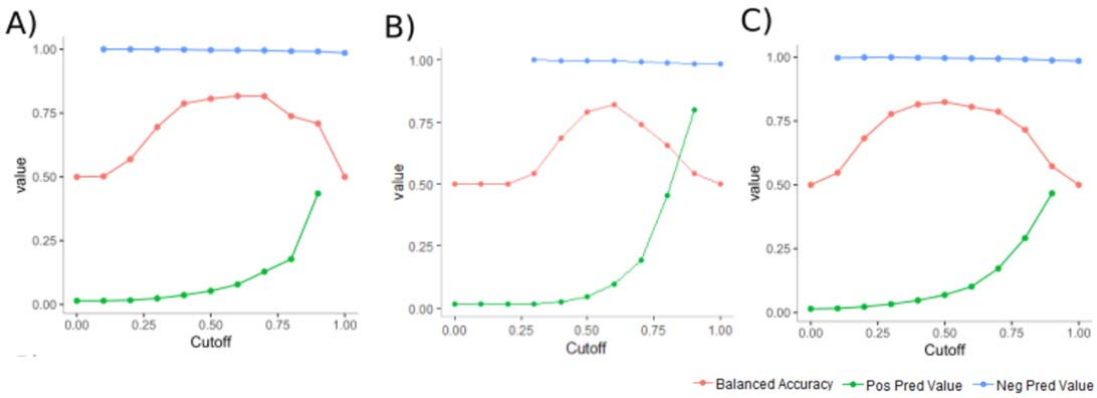

Percentage of missing instances
Ethnicity
GAD
Heavy drinking
OCD
Panic disorder
SAD
Obesity
Negative life events
Self-report health
Non-smoker
Marital status
Antidepressants
Benzodiazepines
Familial income
University degree
Age
Sex
0
0.2
0.4
0.6
0.8
1
1.2
1.4

Figure S2: PPVs and NPVs values for different cut-offs of class boundaries.



7. CONSIDERAÇÕES FINAIS

Esta tese de doutorado teve como resultado a produção de três artigos: uma revisão sistemática e metanálise, publicada em setembro de 2017; uma segunda revisão sistemática, que está sendo reescrita conforme sugestão dos revisores para nova submissão; e um terceiro artigo, submetido e atualmente em revisão.

No primeiro artigo, discutimos as múltiplas aplicações de técnicas de *machine learning* no contexto do transtorno de humor bipolar. Nós mostramos que técnicas de neuroimagem podem ajudar a diferenciar entre transtornos psiquiátricos. Por exemplo, diferenciar entre depressão unipolar e bipolar teria um grande impacto no que diz respeito ao tratamento e prognóstico destes transtornos. Ademais, a predição do tratamento mais adequado para cada indivíduo permitiria uma recuperação mais rápida dos episódios de humor e, consequentemente, menos prejuízos aos pacientes (25,26). Além disso, a predição de desfechos desfavoráveis, como tentativas de suicídio, hospitalizações e novos episódios de humor, pode ser integral para o desenvolvimento de uma psiquiatria baseada em prevenção e com intervenções moldadas de acordo com as necessidades e características de cada indivíduo. Por último, discutimos como o uso de técnicas não supervisionadas de *machine learning* podem ser essenciais para a descoberta de novos fenótipos de transtorno de humor bipolar que possam ser mais relevantes em relação à trajetória e prognóstico do transtorno.

No segundo artigo, nós discutimos o uso de *machine learning* para prever resposta em estudos com intervenções farmacológicas e não-farmacológicas. Nós discutimos problemas atuais dos ensaios clínicos, como o fato destes apenas produzirem resultados a nível de grupo, e não incluírem amostras representativas da heterogeneidade dos transtornos psiquiátricos (27). Nós identificamos 61 estudos que utilizaram diferentes níveis de dados para predição de resposta a tratamento: dados clínicos e sociodemográficos, marcadores séricos, EEG e neuroimagem. Os estudos usando EEG e neuroimagem oferecem vantagens, não só em termos de melhor performance dos modelos, mas também por produzirem marcadores mais objetivos. Dados clínicos, isoladamente ou em adição a outros níveis de dados, mostraram resultados heterogêneos, inclusive com piora da performance em alguns modelos que combinaram estes dados com biomarcadores. Por fim, discutimos limitações metodológicas destes estudos e recomendações sobre o uso de *machine learning* em ensaios clínicos. Novos

estudos deveriam ser desenvolvidos com foco em *big data* e *machine learning*, incluindo grandes amostras, e com critérios de exclusão e inclusão mais flexíveis. Preferencialmente, estes ensaios deveriam ser multicêntricos, para possibilitar adequados teste e validação dos modelos preditivos em diferentes amostras (27).

Finalmente, no terceiro artigo, nós desenvolvemos um modelo para prever o curso do transtorno depressivo, incluindo incidência e persistência de depressão em uma grande coorte ocupacional (ELSA-Brasil) (28). Nós demonstramos que é possível predizer com excelente performance a presença, persistência e incidência de depressão. Além disso, também mostramos que os modelos conseguem diferenciar pacientes com depressão daqueles sem depressão utilizando dados clínicos e sociodemográficos. Em especial, a presença de outras comorbidades psiquiátricas, como transtorno de ansiedade generalizada e transtorno obsessivo-compulsivo, estão entre as variáveis mais relevantes deste modelo preditivo.

O uso de *machine learning* na literatura ainda possui limitações e problemas metodológicos. A maior parte dos estudos usa bancos de dados previamente desenvolvidos para outros fins, e poucos estudos foram desenhados especificamente para o uso de *machine learning* e *big data*. Boa parte dos estudos ainda usa amostras pequenas ou com um reduzido número de variáveis, enquanto a melhor aplicação desta técnica seria com o maior número de pacientes e variáveis possível. Outro problema é o nível de interferência nos modelos, com as variáveis sendo manipuladas de diversas formas pelos autores antes da análise, em detrimento de abordagens *data-driven*. A falta de estudos validando estes modelos prospectivamente é outra importante limitação. Poucos estudos parecem ter ido além do ambiente de pesquisa e tentado aplicar estes modelos como ferramentas clínicas. Em relação ao diagnóstico psiquiátrico, há que se ressaltar que as atuais categorias diagnósticas são definidas apenas por sinais e sintomas, e que seus critérios foram definidos por especialistas (13). Treinar modelos de *machine learning* para prever diagnósticos que foram definidos ignorando a complexidade neurofisiológica dos transtornos é uma limitação em si só. O uso de outros marcadores, como neuroimagem, eletroencefalograma, genética e etc., pode propiciar a descoberta de novas categorias diagnósticas que incluam características patofisiológicas dos transtornos e tenham maior validade para prever a trajetória destas doenças.

Não obstante estas limitações, o uso da tecnologia deve se tornar cada vez mais ubíquo em nosso dia-a-dia, com oportunidades ilimitadas para melhorar os sistemas de saúde e a prática clínica. O fenótipo digital, por exemplo, termo cunhado por Jukka-Pekka Onnela para se referir à quantificação da interação individual entre uma pessoa e seus dispositivos eletrônicos, carrega um enorme potencial para a aquisição de dados e a tomada de decisões em tempo real. Avanços em técnicas de neuroimagem, eletrofisiologia, sensores vestíveis, dentre outros, oportunizam a busca por marcadores mais objetivos para caracterizar os transtornos psiquiátricos e colaborar para um manejo mais adequado e preciso destes. No centro desta revolução, impulsionado pelo exponencial avanço na capacidade de processamento e efetividade dos computadores, está o campo de *machine learning*, capaz de analisar esses novos complexos sistemas de dados. Em adição, o próprio campo de *machine learning* apresenta avanços exponenciais, com técnicas como aprendizado profundo, aprendizado por reforço, *ensemble learning*, redes antagônicas geradoras, e neuroevolução sendo aprimoradas e aplicadas para as mais diversas funções. O desenvolvimento de sistemas e processos tecnológicos de crescente complexidade, e a constante evolução e aprimoramento de técnicas avançadas que possam interpretar estes sistemas e processos, são dois fenômenos sinergéticos que, em última análise, serão promotores da maior revolução já vista não só em psiquiatria em saúde mental, mas na ciência como um todo.

Esta tese traz como contribuição a relevância e o impacto que as técnicas de *machine learning* podem trazer à psiquiatria e à saúde mental, em especial para os sistemas de saúde e os pacientes. Muitos dos desafios atualmente presentes nessas áreas podem ser solucionados com o uso destas técnicas, promovendo uma psiquiatria preditiva e personalizada, centrada nas características e necessidades de cada indivíduo.

8. REFERÊNCIAS BIBLIOGRÁFICAS

1. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ [Internet]. 1996 Jan 13;312(7023):71–2. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.312.7023.71
2. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med [Internet]. 2016 Aug;21(4):125–7. Available from: http://ebm.bmj.com/%0Ahttp://www.isehc.net/wp-content/uploads/2011/12/October-2015.pdf
3. Greenhalgh T, Howick J, Maskrey N. Evidence based medicine: a movement in crisis? BMJ [Internet]. 2014 Jun 13;348(jun13 4):g3725–g3725. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.g3725
4. Martino DJ, Strejilevich SA, Scapola M, Igoa A, Marengo E, Ais ED, et al. Heterogeneity in cognitive functioning among patients with bipolar disorder. J Affect Disord. 2008 Jul;109(1–2):149–56.
5. Modabbernia A, Taslimi S, Brietzke E, Ashrafi M. Cytokine Alterations in Bipolar Disorder: A Meta-Analysis of 30 Studies. Biol Psychiatry [Internet]. 2013 Jul;74(1):15–25. Available from: http://dx.doi.org/10.1016/j.biopsych.2013.01.007
6. Lima F, Rabelo-da-ponte FD, Bücker J, Czepielewski L, Hasse-sousa M, Telesca R, et al. Identifying cognitive subgroups in bipolar disorder : A cluster analysis. J Affect Disord [Internet]. 2019;246(August 2018):252–61. Available from: https://doi.org/10.1016/j.jad.2018.12.044
7. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. The Lancet Psychiatry [Internet]. 2016 Jan;3(1):13–5. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2215036615005490
8. Khoury MJ, Ioannidis JPA. Big data meets public health. Science (80-) [Internet]. 2014 Nov 28;346(6213):1054–5. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.aaa2709
9. Monteith S, Glenn T, Geddes J, Bauer M. Big data are coming to psychiatry: a general introduction. Int J Bipolar Disord [Internet]. 2015;3(1):21. Available from: http://europepmc.org/abstract/MED/26440506
10. McIntosh AM, Stewart R, John A, Smith DJ, Davis K, Sudlow C, et al. Data science for mental health: a UK perspective on a global challenge. The Lancet Psychiatry [Internet]. 2016;3(10):993–8. Available from: http://dx.doi.org/10.1016/S2215-0366(16)30089-X
11. Lantz B. Machine Learning with R - Second Edition [Internet]. Intergovernmental Panel on Climate Change, editor. Packt Publishing. Cambridge: Cambridge University Press; 2015. Available from:

http://books.google.com/books?id=ZQu8AQAAQBAJ&printsec=frontcover&dq=intitle:Machine+Learning+with+R&hl=&cd=1&source=gbs_api%5Cnpapers2://publication/uuid/46164A51-A282-4F67-8397-9FA79F39B5B7

12. Bzdok PD, Meyer-Lindenberg PA. Machine learning for precision psychiatry. :1–16.
13. Insel TR, Cuthbert BN. Brain disorders? Precisely. Science (80-) [Internet]. 2015 May 1;348(6234):499–500. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.aab2358
14. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016 Feb;19(3):404–13.
15. Passos IC, Mwangi B. Machine learning-guided intervention trials to predict treatment response at an individual patient level: an important second step following randomized clinical trials. Mol Psychiatry [Internet]. 2018;1–2. Available from: http://www.nature.com/articles/s41380-018-0250-y
16. Kapczinski NS, Mwangi B, Cassidy RM, Librenza-Garcia D, Bermudez MB, Kauer-Sant'anna M, et al. Neuroprogression and illness trajectories in bipolar disorder. Expert Rev Neurother. 2017;17(3).
17. Wu MJ, Mwangi B, Bauer IE, Passos IC, Sanches M, Zunta-Soares GB, et al. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. Neuroimage [Internet]. 2017;145:254–64. Available from: http://dx.doi.org/10.1016/j.neuroimage.2016.02.016
18. Fusar-Poli P, Solmi M, Brondino N, Davies C, Chae C, Politi P, et al. Transdiagnostic psychiatry: a systematic review. World Psychiatry. 2019;18(2):192–207.
19. De Almeida JRC, Phillips ML. Distinguishing between unipolar depression and bipolar depression: Current and future clinical and neuroimaging perspectives. Biol Psychiatry [Internet]. 2013;73(2):111–8. Available from: http://dx.doi.org/10.1016/j.biopsych.2012.06.010
20. Whitfield-Gabrieli S, Ghosh SS, Nieto-Castanon A, Saygin Z, Doehrmann O, Chai XJ, et al. Brain connectomics predict response to treatment in social anxiety disorder. Mol Psychiatry [Internet]. 2015;(January):1–6. Available from: http://www.nature.com/doifinder/10.1038/mp.2015.109
21. Hahn T, Kircher T, Straube B, Wittchen H-U, Konrad C, Ströhle A, et al. Predicting Treatment Response to Cognitive Behavioral Therapy in Panic Disorder With Agoraphobia by Integrating Local Neural Information. JAMA Psychiatry [Internet]. 2015;72(1):68–74. Available from: http://archpsyc.jamanetwork.com/article.aspx?articleID=1936093&utm_source=Silverchair Information Systems&utm_medium=email&utm_campaign=JAMAPsychiatry:NewIssue01/07/2015%5

Cnhttp://archpsyc.jamanetwork.com/article.aspx?articleID=1936093&utm_source=Silverchai

22. Redlich R, Opel N, Grotegerd D, Dohm K, Zaremba D, Bürger C, et al. Prediction of Individual Response to Electroconvulsive Therapy via Machine Learning on Structural Magnetic Resonance Imaging Data. JAMA Psychiatry [Internet]. 2016 Jun 1;73(6):557. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27145449
23. Salagre E, Dodd S, Aedo A, Rosa A, Amoretti S. Toward Precision Psychiatry in Bipolar Disorder : Staging 2.0. 2018;9(November).
24. Nasrallah HA. The dawn of precision psychiatry. 16(12):16–8.
25. Passos IC, Mwangi B, Vieta E, Berk M, Kapczinski F. Areas of controversy in neuroprogression in bipolar disorder. Acta Psychiatr Scand. 2016 Apr;
26. Frey BN, Zunta-Soares GB, Caetano SC, Nicoletti MA, Hatch JP, Brambilla P, et al. Illness duration and total brain gray matter in bipolar disorder: Evidence for neurodegeneration? Eur Neuropsychopharmacol. 2008;18(10):717–22.
27. Passos IC, Mwangi B. Machine learning-guided intervention trials to predict treatment response at an individual patient level: an important second step following randomized clinical trials. Mol Psychiatry [Internet]. 2018 Sep 21; Available from: http://www.nature.com/articles/s41380-018-0250-y
28. Aquino EML, Barreto SM, Bensenor IM, Carvalho MS, Chor D, Duncan BB, et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): Objectives and Design. Am J Epidemiol [Internet]. 2012 Feb 15;175(4):315–24. Available from: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwr294

8. ANEXOS

8.1. Anexo 1

Posters, resumos de congresso e apresentações orais produzidos durante o período que compreendeu este doutorado:

1. Pinto JV, **Librenza-Garcia D**, Baldez DP, Rosa AR, Kapczinski F, Kauer-Sant'Anna M, Passos IC. Predicting functional impairment in bipolar disorder: a pilot study with a machine learning approach. Poster session, *European Congress of Neuropsychopharmacology*, **2017**.
2. **Librenza-Garcia D**, Noll G, Przybylski L, Costanzi M, de Azevedo Cardoso T, Kapczisnki F, Passos IC. Rapid cycling prediction in bipolar disorder using clinical variables: A machine learning, proof-of-concept study. Poster session, *The International Society of Bipolar Disorders (ISBD) conference*, **2018**. Doi: 10.1111/bdi.12619
3. de Aguiar BW, **Librenza-Garcia D**, Spanemberg L, Caldieraro MA, Watts D, Boctor M, Fleck M, Passos IC, Kapczinski F. Differential biomarker signatures in unipolar and bipolar depression: A machine learning approach. Poster session, *The International Society of Bipolar Disorders (ISBD) conference,* **2018**. Doi: 10.1111/bdi.12619
4. Pinto JV, **Librenza-Garcia D**, Przybylski L, Noll G, Kauer Sant'Anna M, Ribeiro Rosa A, Kapczinski F, Passos IC. Cognitive functioning impairment prediction in patients with bipolar disorder: a pilot study using machine learning techniques. Poster session, *The International Society of Bipolar Disorders (ISBD) conference*, **2018**. Doi: 10.1111/bdi.12619

Apresentação oral:

1. Acosta JR, **Librenza-Garcia** D, Zortéa F, Tramontina S, Passos IC. Predicting lifetime psychosis in bipolar disorder youths using machine learning techniques. Rapid communication session, *The International Society of Bipolar Disorders (ISBD) conference*, **2018**. Doi: 10.1111/bdi.12618

8.2. Anexo 2

Artigos originais e capítulos de livros produzidos durante o período que compreendeu este doutorado.

Artigos publicados:

1. Ramos BR, **Librenza-Garcia D**, Zortea F, Watts D, Zeni CP, Tramontina S, Passos IC. Clinical Differences Between Patients With Pediatric Bipolar Disorder With And Without A Parental History Of Bipolar Disorder. *Psychiatry Res*. **2019**. In press. DOI: 10.1016/j.psychres.2019.112501
2. Tietbohl-Santos B, Chiamenti P, **Librenza-Garcia D**, Cassidy R, Zimerman A, Manfro GG, Kapczinski F, Passos IC. Risk factors for suicidality in patients with panic disorder: A systematic review and meta-analysis. *Neurosci Biobehav Rev*. **2019** Jul 31;105:34-38. doi: 10.1016/j.neubiorev.2019.07.022.
3. de Ávila Berni G, Rabelo-da-Ponte FD, **Librenza-Garcia D**, Boeira MV, Kauer-Sant'Anna M, Passos IC, Kapczinski F. Potential use of text classification tools as signatures of suicidal behavior: A proof-of-concept study using Virginia Woolf's personal writings. *PLoS One*. **2018** Oct 24;13(10):e0204820. Doi: 10.1371/journal.pone.0204820
4. Pinto JV, Passos IC, **Librenza-Garcia D**, Marcon G, Schneider MA, Conte JH, da Silva JPA, Lima LP, Quincozes-Santos A, Kauer-Sant Anna M, Kapczinski F. Neuron-glia Interaction as a Possible Pathophysiological Mechanism of Bipolar Disorder. *Curr Neuropharmacol*. **2018**;16(5):519-532. Doi: 10.2174/1570159X15666170828170921
5. Kapczinski NS, Mwangi B, Cassidy RM, **Librenza-Garcia D**, Bermudez MB, Kauer-Sant'anna M, Kapczinski F, Passos IC. Neuroprogression and illness trajectories in bipolar disorder. *Expert Rev Neurother*. **2017** Mar;17(3):277-285. doi: 10.1080/14737175.2017.1240615.

Artigos aceitos:

1. Acosta JR, Zortéa F, **Librenza-Garcia D**, Watts D, Francisco AP, Raffa B, Motta GLC, Kohmann Andre, Mugnol FE, Tramontina S, Passos IC. Bullying and psychotic symptoms in youth with bipolar disorder. *Journal of affective disorders.* **2019.**

2. Passos IC, Ballester P, Barros R, **Librenza-Garcia D**, Mwangi B, Birmaher B, Brietzke E, Hajek T, López-Jaramillo C, Mansur R, Alda M, Haarman BCM, Isometsa E, Lam R, McIntyre R, Minuzzi L, Kessing L, Yatham L, Duffy A, Kapczinski F. Machine learning and big data analytics in bipolar disorder: A Position paper from the International Society for Bipolar Disorders (ISBD) Big Data Task Force. *Bipolar Disorders.* **2019.**

Artigos submetidos:

1. Wollenhaupt-Aguiar B, **Librenza-Garcia D**, Bristot G, Przybylski L, Stertz L, Kubiachi Burque R, Mendes Ceresér K, Spanemberg L, Caldieraro M, Frey B, Fleck M, Kauer-Sant'Anna M, Passos IC, Kapczinski F. Differential Biomarker Signatures In Unipolar And Bipolar Depression: A Machine Learning Approach. Submitted to *Australian and New Zealand Journal of Psychiatry*.
2. Reboucas DB, Sartori JM, **Librenza-Garcia D**, Massuda R, Czepielewski LS, Passos IC, Gama C. Accelerated aging signatures in individuals with schizophrenia and their unaffected siblings: a machine learning approach. Submitted to *Schizophrenia Research*.

Capítulo de livro:

1. Chapter 9: Ethics in the Era of Big Data. Personalized psychiatry: Big Data Analytics in Mental Health. **Librenza-Garcia, D**. Springer, **2019**. Doi: 10.1007/978-3-030-03553-2_9.