DR. IVES CAVALCANTE PASSOS (Orcid ID : 0000-0001-6407-8219)
DR. ELISA BRIETZKE (Orcid ID : 0000-0003-2697-1342)
DR. TOMAS HAJEK (Orcid ID : 0000-0003-0281-8458)
DR. MARTIN ALDA (Orcid ID : 0000-0001-9544-3944)
MR. BARTHOLOMEUS C.M. HAARMAN (Orcid ID : 0000-0002-9006-8863)
PROF. ROGER S MCINTYRE (Orcid ID : 0000-0003-4733-2523)
DR. LARS KESSING (Orcid ID : 0000-0001-9377-9436)
PROF. ANNE DUFFY (Orcid ID : 0000-0002-5895-075X)
DR. FLAVIO KAPCZINSKI (Orcid ID : 0000-0001-8738-856X)

Article type : Review

TITLE PAGE

Machine learning and big data analytics in bipolar disorder: A Position paper from the International Society for Bipolar Disorders (ISBD) Big Data Task Force

Authors: Ives Cavalcante Passos, MD, PhD[1]; Pedro Ballester[2]; Rodrigo Coelho Barros, PhD[2]; Diego Librenza-Garcia, MD[3]; Benson Mwangi, PhD[4]; Boris Birmaher, MD[5]; Elisa Brietzke, MD, PhD[6]; Tomas Hajek, MD, PhD [7,10]; Carlos Lopez Jaramillo, MD, MSc, PhD[8]; Rodrigo B. Mansur, MD, PhD[9]; Martin Alda, MD[7]; Bartholomeus C.M. ('Benno') Haarman, MD[11]; Erkki Isometsa, MD, PhD[12]; Raymond W Lam, MD[13]; MD; Roger S. McIntyre, MD, FRCPC[14]; Luciano Minuzzi, MD, PhD[3]; Lars Vedel Kessing, MD, DMSc[15]; Lakshmi N. Yatham, MBBS[13]; Anne Duffy, MD, MSc [6]; Flavio Kapczinski MSc, MD, PhD, FRCPC[3].

1. Laboratory of Molecular Psychiatry and Bipolar Disorder Program, 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, RS, Brazil.
2. School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Rio Grande do Sul, RS, Brazil.
3. Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, ON, Canada.
4. UT Center of Excellence on Mood Disorders, Department of Psychiatry and Behavioral Sciences, The University of Texas Health Science Center at Houston, McGovern Medical School, Houston, TX, USA.

5. Department of Psychiatry, Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania.
6. Department of Psychiatry, Queen’s University, Kingston, ON, Canada.
7. Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, Canada.
8. Research Group in Psychiatry, Department of Psychiatry, Faculty of Medicine, University of Antioquia; Mood Disorders Program, Hospital Universitario San Vicente Fundación, Medellín, Colombia.
9. Mood Disorders Psychopharmacology Unit (MDPU), University Health Network, University of Toronto, Toronto, Canada.
10. National Institute of Mental Health, Klecany, Czech Republic
11. Department of Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
12. Department of Psychiatry, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
13. Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada.
14. Department of Psychiatry, University of Toronto, Toronto, ON, Canada
15. Copenhagen Affective Disorder Research Center (CADIC), Psychiatric Center Copenhagen, Copenhagen University Hospital, Denmark.

Corresponding author:
Flávio Kapczinski, MD, PhD
Full Professor at McMaster University
Director of MINDs – Graduation Program in Mental Health (McMaster University)
Email: flavio.kapczinski@gmail.com

Email for authors:
Ives Cavalcante Passos: ivescp1@gmail.com
Pedro Ballester: pedro.ballester@acad.pucrs.br
Rodrigo Coelho Barros: rodrigo.barros@pucrs.br
Diego Librenza-Garcia: diegolibrenzagarcia@gmail.com
Benson Mwangi: benson.mwangi@gmail.com
Boris Birmaher: BirmaherB@upmc.edu
Elisa Brietzke: elisabrietzke@hotmail.com
Tomas Hajek: Tomas.Hajek@dal.ca
Carlos Lopez Jaramillo: carloslopezjaramillo@gmail.com
Rodrigo Mansur: rodrigo.mansur@uhn.ca

Martin Alda: malda@dal.ca

Bartholomeus C.M. ('Benno') Haarman: b.c.m.haarman@rug.nl

Erkki Isometsa: erkki.isometsa@helsinki.fi

Raymond W Lam: r.lam@ubc.ca

Roger S McIntyre: roger.mcintyre@uhn.ca

Luciano Minuzzi: lminuzzi@gmail.com

Lars Vedel Kessing: lars.vedel.kessing@regionh.dk

Lakshmi Yatham: l.yatham@ubc.ca

Anne Duffy: anne.duffy@queensu.ca

Flávio Kapczinski: flavio.kapczinski@gmail.com

Abstract

Objectives: The International Society for Bipolar Disorders (ISBD) Big Data Task Force assembled leading researchers in the field of bipolar disorder (BD), machine learning, and big data with extensive experience to evaluate the rationale of machine learning and big data analytics strategies for BD.

Method: A task force was convened to examine and integrate findings from the scientific literature related to machine learning and big data based studies to clarify terminology and to describe challenges and potential applications in the field of BD. We also systematically searched PubMed, Embase, and Web of Science for articles published up to January 2019 that used machine learning in BD.

Results: The results suggested that big data analytics has the potential to provide risk calculators to aid in treatment decisions and predict clinical prognosis, including suicidality, for individual patients. This approach can advance diagnosis by enabling discovery of more relevant data-driven phenotypes, as well as by predicting transition to the disorder in high-risk unaffected subjects. We also discuss the most frequent challenges that big data analytics applications can face, such as heterogeneity, lack of external validation and replication of some studies, cost and non-stationary distribution of the data, and lack of appropriate funding.

Conclusion: Machine learning based studies, including atheoretical data-driven big data approaches, provide an opportunity to more accurately detect those who are at risk, parse-relevant phenotypes as well as inform treatment selection and prognosis. However, several methodological challenges need to be addressed in order to translate research findings to clinical settings.

Keywords: bipolar disorder, big data, machine learning, deep learning, data mining, personalized psychiatry, risk prediction, predictive psychiatry.

Introduction

Bipolar disorder (BD) has a worldwide prevalence of about 2% with subclinical variants affecting another 2% of the population.[1] According to the World Health Organization, BD is among the top 10 leading causes of disability-adjusted life years in young adults.[2] Rates of completed suicide in patients with BD are 7.8% in men and 4.9% in women.[3] These patients commonly endure prolonged periods of trial and error before an effective treatment among the possible options is found. Although interventions to treat and prevent mood episodes are available and detailed in Guidelines,[4,5] unselected treatment or guideline drive treatment is frequently suboptimal, and about 60% of the patients relapse into depression or mania within two years of treatment initiation.[6,7] Early intervention is critical in BD to prevent progression and complications such as suicide attempts,[8–10] however, current approaches to diagnosing BD leave room for improvement, since there is an average delay of ten years between the first symptoms and a formal diagnosis.[11]

Randomized clinical trials (RCTs) and meta-analyses have helped to find effective treatments for BD, such as lamotrigine[12] and quetiapine,[13] by using traditional statistical methods, which primarily provide average group-level results based on measures of central tendency and variance. This approach allows us to make broad generalizations about patients with BD in regard to specific treatments. However, it fails to detect nuances related to an individual patient, and significant results may not represent a real benefit for individuals.[14] Indeed, subjects included in clinical trials do not consistently reflect patients with BD from real-world clinical scenarios – in fact, the very idiosyncrasies that characterize most of these patients, such as the multi-morbidity profiles, are often exclusion criteria in clinical trials.[15] In addition, evidence suggests that BD is a heterogeneous disorder with valid subgroups, each with a specific responsiveness to prophylactic treatment.[16,17] Big data analysis by machine learning techniques provides the means to move beyond group level statistics into individual subject classification based on accuracy, sensitivity, specificity, and area under the ROC curve (AUC).

Another elusive goal in the current study of BD is the prediction of prognosis in individuals already affected and transition to full-blown illness in those ones at risk.[18] The linear association between risk factors and clinical outcomes is important to understand BD,[19] however, they do not objectively stratify which subjects will develop BD or, when affected, what will be their prognosis.[20,21] The integration of a huge number of risk factors necessarily requires new analytical tools. To fill these gaps, big data analytics is being used in psychiatry to provide predictive models for both clinical practice and public health systems.[22]

In this manuscript, the Big Data Task Force of the International Society for Bipolar Disorder (ISBD) will explore the role of machine learning techniques and big data in improving outcomes prediction in prevention, diagnosis, and treatment of individuals with BD. Specifically, we will 1) define big data and machine learning techniques and outline the issues that need to be considered in machine-learning-based studies; 2) update a systematic review of published studies on machine learning and big data in BD to illustrate where the field is at right now; and 3) identify the obstacles for application of these methodologies in BD and propose strategies to overcome them.

Methods

A Task Force was convened to examine, discuss, and integrate findings from the scientific literature related to machine learning based studies and big data to clarify terminology and to describe challenges and potential applications in the field of BD. We also updated a systematic review published by our group.[23]

Search strategy

For the systematic review, we searched PubMed, Embase, and Web of Science for articles published between January 1960, and January 2019 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.

Eligibility criteria

This systematic review was performed according to the PRISMA statement.[24] Articles met the inclusion criteria if they assessed patients with BD using machine learning techniques. Technical and theoretical studies that used machine learning techniques but did not assess patients with BD were excluded. We also excluded studies that included only individuals below 18 years of age.

Data collection, extraction, and statistical analysis

Two researchers (DLG and PB) 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.

Definitions

Big Data

The first definition of big data focuses on the 3 Vs - velocity, volume, and variety. “Velocity” refers to the speed at which the data is generated, while “Volume” refers to the amount of data, and is readily demonstrated by for example the storage space needed. “Variety” refers to the diverse nature of data collected from many sources. For healthcare, this means that data for understanding one's behaviour should not be collected only from anamnesis, exams, and clinical questionnaires. Instead, data should be pervasive and gathered in multiple modalities, including patient behavior and social relationships. A more recent definition adds *veracity* and *value* as two additional Vs.[25] “Veracity” concerns whether or not we can trust the data we gather, and “Value” refers to the fact that we must integrate all of the aforementioned pillars of big data towards improvement on how we treat and monitor patients, thus generating value for families, caregivers, and patients suffering from the BD.[26]

One example of a big data application is ecological momentary assessment, which refers to the continuous collection of data by smartphones or personal devices. Its potential is based on the assumption that traditional clinical approaches in assessment of mood symptoms are unsatisfactory since they require that the patient summarize their symptoms over a defined time framework (e.g. one month) in one sentence: Over the past month I felt: ”good”, “not so good”, “very depressed”, or “a little bit manic”. Therefore, there is no granularity in mood reporting and no reliable information about variability or association with other symptoms or exposures, which would be putatively more in line with neurobiology. Additionally, there is no ecological validity to this kind of reporting (patients behave differently when they are in our offices). We need a comprehensive, ecologically valid, precise and passive collection of data as proposed in some studies.[27,28] However, such rich data collection will generate millions of data points that will require specific approaches for analysis.

Machine Learning

Most of the process for finding useful patterns in data that have translational meaning and can be incorporated in day-to-day practice is possible through machine learning approaches - a powerful tool for pattern recognition and responsible for most of the recent advances in artificial intelligence. Through almost no pre-assumptions and a nonlinear function canvas, we can model complex patterns that can identify relationships between large amounts of and diverse data.[29] This change in perspective introduces more flexibility in our groups (we may include fewer constraints in inclusion/exclusion criteria from clinical trials), while providing important information on a clinical outcome by taking into account heterogeneity. Furthermore, by incorporating feature selection in the process, we can automatically select subgroups of predictors that are most relevant for a model, providing simpler and more clinically useful results.[30]

But how does machine learning operate? Usually, machines receive data from a certain scenario, ranging from simple online surveys responses to complex biomarkers measurements, such as genetics or neuroimaging, and approximate a function that best fits the predictors.[31] This process is called training and it represents the process of *learning*. In the context of healthcare, the learning method is usually contained in one of two paradigms, *supervised* or *unsupervised* learning. In supervised learning, the user feeds the machine with predictors and expected outcome. The machine thus learns a mapping $X \rightarrow Y$ from the predictor space X to the outcome space Y. This paradigm includes tasks such as predicting suicide attempts in patients with mood disorders based on prior clinical or demographic variables.[32] However, unsupervised learning does not depend on Y and has clustering as its most common class of algorithms. Clustering can find hidden groups underlying the predictors' variance and help users explain phenomena. This type of learning includes finding subgroups of patients that share underlying characteristics, such as suicidaity or neurocognitive impairment in a proportion of patients with BD.[33]

To make easier the transition for practitioners to the lingo introduced in this new field, we provide a quick terminology reference in Table 1. Additionally, Table 2 provides important points to be considered in machine-learning-based studies and Figure 1 shows how a machine learning experiment should be conducted. It is important to note that machine learning is but one of several methods that can be used to analyze big data. For instance, discriminant analysis or various methods of principal component analysis, cluster analysis, factor analysis can all be used with various assumptions being met for analyzing large data sets and for differentiating either pre-defined groups ("supervised") or hypothesis-free (data-generated) subgroups ("unsupervised"). We chose machine learning techniques in the present review because of its ability to model complex patterns, including non-linear relationships.

How will machine learning and big data analytics contribute to the field of bipolar disorder?

We found 1124 potential abstracts and included 91 articles in the present review, with one of these added after reference screening (Figure S1). We found 37 additional articles compared to the prior systematic review.[23] We briefly described below how machine learning and big data will contribute to the field of BD by highlighting some of the included articles. The most relevant characteristics and findings of each of the 89 included studies are described in the supplemental material (tables S1, S2, S3, and S4).

Diagnostic studies

Structural and functional neuroimaging, as well as diffusion tensor imaging (DTI), have been widely used in classification studies.[34–36] A recent large study applied support vector machines to MRI data (regional cortical thickness, surface area, subcortical volumes) from 853 patients with BD and 2167 control participants from 13 cohorts in the ENIGMA consortium.[37] Authors found an AUC of 0.71 in differentiating BD from controls. Additionally, a recent meta-analysis showed an AUC of the summary ROC curve of 0.70 for structural and of 0.75 for functional neuroimaging studies.[23] We found one study that used DTI and included 67 unmedicated depressed patients, including 31 patients with BD and 36 with major depressive disorder (MDD). Authors found that the fractional anisotropy (FA) tract profile of the left anterior thalamic radiation can be used to differentiate between the BD and MDD patients at an accuracy of 68.33%.[38] Other sets of data, such as genetics,[39–44] electroencephalogram,[45–50] neuropsychological tests,[51,52] blood biomarkers,[28,53–56] text,[57] facial expressions,[58] and speech[59] were also used to classify patients with BD from healthy controls or from other psychiatric disorders (Table S1). It is worth mentioning that a study used the concept of ecological momentary assessment to distinguish patients with BD from patients with borderline personality disorder and healthy controls by using daily mood ratings from a smartphone app.[60] Authors reported that the methodology classified 75% of participants into the correct diagnostic group compared with 54% using standard approaches.

These studies may provide a more objective diagnosis for BD in the near future. However, some limitations should be addressed to allow translation of these findings to the clinical practice. First, the initial AUC of the predictive models should be improved by including other layers of data and applying a multimodal data approach. Second, most of the studies lack external validity since they were built by using only patients with BD and controls. Therefore, population and largely representative studies, including other psychiatric or neuropsychiatric disorders, should be conducted to ensure generalization of the proposed models. Third, we still do not know how these models will perform in face of patients from different stages of the disorder.[8,15,61] In this sense, a recent study found that a machine learning model developed by using the relevance vector machine algorithm and white matter from structural MRI was more accurate in identifying patients with BD at the late stage.[62]

Prediction of poor clinical outcomes

Some studies used machine learning techniques to predict suicidality and mood episode relapse. A study tested a set of machine learning algorithms coupled with clinical and demographic variables to develop a clinical signature of suicidality in 144 patients with mood disorders, including BD.[63] The study reported a balanced accuracy of 72% and an AUC of 0.77 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 in the model. Other studies also predicted suicidality by using machine learning coupled with a combined genomic and clinical risk assessment approach and built models with an AUC of 0.98[64] and 0.82[65] in patients with BD. It is also worth mentioning that a recent text classification study used letters and diaries of Virginia Woolf to identify written patterns associated with suicide.[66] Authors found an AUC of 0.80 and a balanced accuracy of 80.45% by using Naïve-Bayes machine-learning algorithm.

Another study used demographic and clinical features, including follow-up variables, to assess depression relapse in 108 patients with BD and achieved an accuracy of 85%, and a sensitivity of 92%.[67] Furthermore, a study using voice features collected in phone calls to classify patients' affective states, achieved an AUC of 0.78 (depressed vs. euthymic) and 0.89 (manic/mixed vs. euthymic).[68] These proof-of-concept and experimental protocols illustrate machine learning's potential to aid in the clinical assessment of BD patients, yielding models with sufficient accuracy to monitor mood states in real time which may help assess disease activity and advance early intervention. Although promising, most of these studies included small samples, and, therefore, need to be interpreted with caution requiring adequate model validation in different settings and populations. Table S2 in the supplemental material presents studies that used machine learning methods to predict clinical outcomes in BD.

Selection of treatment

The incorporation of tools from machine learning to guide trials for better-tailored interventions is a necessary next step to move beyond current group-based approaches.[69] These models can be displayed as user-friendly calculators, and incorporated into the clinical workflows of electronic medical records.[70] For instance, if a calculator predicts that a given patient is unlikely to respond to an intervention, the clinician could then consider alternatives.[71] Accordingly, patients would benefit from more precise treatment plans with less delay avoiding the associated burden of untreated illness. These calculators estimate the probability of a particular outcome and are ideal for assessing the multimorbidity profile and other nuances found in patients with BD - as long as their heterogeneity is represented in the training dataset.[69] In the field of BD, a pilot study developed a treatment response calculator for lithium.[72] Authors included 20 subjects with first-episode bipolar mania who received lithium over 8 weeks. A machine learning model coupled with fMRI and 1H-MRS scans data at baseline pretreatment was trained and validated. The model was able to predict post-treatment symptom reductions at 8 weeks with 80% accuracy in the validation phase.

Machine learning guided interventions will not only facilitate the selection of treatment based on efficacy but also aid in the prevention of side effects.[69] In this sense, a study with more than 5700 patients undergoing lithium treatment built a predictive algorithm to renal insufficiency by using logistic regression and electronic medical records.[73] Authors found an AUC of 0.81 in an independent testing set. Use of lithium more than once daily, lithium levels greater than 0.6mEq/l, and the use of first-generation antipsychotics were independently associated with risk. These findings suggest that risk stratification can be expanded to other treatments and interventions. Moreover, estimating the risk of certain side effects could allow more informed decisions and facilitate the development of prevention strategies.

Finally, machine learning guided trials may have a different design compared to RCTs.[69] First, it is not necessary to have a control group since the aim is to stratify the already known evidence from RCTs among patients with BD. Second, inclusion criteria should not be restrictive since the heterogeneity, multimorbidity profile and other nuances found in patients with BD from real-world clinical scenarios should be represented in the sample. These types of trials will shift the focus from group-level averages to individuals and will ultimately leverage each person's unique clinical and biological profile to improve selection of treatment. Table S3 shows machine learning studies predicting treatment response and adverse effects.[72–76]

Prediction of transition to bipolar disorder

Another elusive goal in the field of BD is the prediction of transition to full-blown illness and its prognosis. The risk to first degree relatives of a patient suffering from BD is estimated at 10-fold that of the general population.[77] Additionally, several risk factors for bipolar and related mood disorders have been identified in those at confirmed familial risk including the presence of subthreshold mood symptoms (hypomanic, depressive and anxiety symptoms),[61,78–80] antecedent non-mood childhood diagnoses (i.e., anxiety and sleep disorders)[81], experiences of childhood abuse and neglect,[82] increased exposure to unstable parental BD[83] and temperament factors.[84] However, the relative contribution of these independent risk factors and the interaction among them in predicting mood disorders is unclear. In addition, it is unlikely that a single biomarker can predict who will develop BD since multiple and complex bio-psychosocial pathways lead to these disorders.[85] However, no study has used machine learning techniques to predict conversion to BD in those at identified risk as yet. Two studies have built risk calculators for illness development using other mathematical approaches. One of these assessed the probability that an offspring of a parent with BD will develop a new-onset bipolar spectrum disorder within the next 5 years.[79] Authors found an AUC of 0.76 by using Cox proportional hazards regression. Another similar study built a risk calculator to predict the individual risk of transition from subthreshold bipolar symptoms to bipolar I or II in youth and reported an AUC of 0.71.[86]

Data clustering using unsupervised and semi-supervised machine learning

BD and other psychiatric disorders are all extremely heterogeneous, in terms of their clinical presentation (which we refer to as 'clinical heterogeneity'), underlying biological causes ('biological heterogeneity') and environmental exposures ('environmental heterogeneity'). These sources of heterogeneity remain a substantial barrier to better understanding the causative mechanisms of psychiatric disorders and to developing optimal treatments and diagnostic tools. It has long been recognized that we must look beyond simple case-control comparisons to be able to deconstruct the heterogeneous phenotype of BD.[87] Unsupervised and semi-supervised machine learning techniques may aid to digest the heterogeneity; however, few studies used these techniques in the field of BD (Table S4).[88–92] In addition, replication remains a major challenge for these approaches. Indeed, none of the studies we have reviewed have been independently replicated.

Obstacles and ethical issues

Although we have a plethora of studies using machine learning and big data approaches to tackle complex questions in BD, knowledge translation to clinical practice is still under-developed[93]. Obstacles, including model validation, computational power, multimodality, assessment of rare events, cost and non-stationary distribution of the data, heterogeneity both phenotypically and etiologically, phenomenological diagnosis, lack of a uniform pipeline for machine learning studies, lack of appropriate funding, and lack of interpretability, need to be addressed.

Model Validation

When training a model, machines can either fit the training data incredibly well or not find a function that suits the data properly. In the former, we should be careful whether the machine is not performing overfitting. In that case, when performing inferences on new data (unseen during the training process), the machine will probably lead to suboptimal results. In the latter case, when the machine underfits the data, it can be easily seen during training. Most of the time we refer to such challenge as a variance and bias tradeoff. Essentially, the goal is to develop a model with enough variance to model complex shapes but not too much to overfit the data.

An additional point related to *overfitting* is that it can also lead to false conclusions on data behavior. By observing the variables that contribute to the result, the researcher could wrongfully declare a new finding for an outcome not able to be replicated in another trial. For that reason, one must follow a protocol to ensure that the findings are robust. There are several approaches here, including *Bootstrap*, *Cross-Validation*, and *Holdout.*[94] The gold standard is debatable but usually consists of performing *Cross-Validation* to find the best fitting model in the training data and finally evaluating once in a *Holdout* set. The *Holdout* set should be totally unseen during the process and ultimately should be collected on an entirely new sample at a different institution when possible.

Computational Power and Quality of Data

There is a clear tradeoff between how much data you have and the quality of the models you can develop. Usually, the more data you gather, the more your model will generalize for unseen instances. Additionally, you lessen the chances of overfitting and it gets easier to properly use the validation protocols. Two other facets of machine learning research in healthcare are challenging. First, collecting large amount of data can be costly and often associated with logistical challenges. Second, the computational power required to address large databases grows exponentially with the data complexity/size. Performing proper *feature selection* and *hyperparameter optimization* in big data can also be challenging.[30] Some studies on unstructured data such as images, text, video, and others, can also approach the problem of using deep learning. Convolutional neural networks, the standard framework for systems in this kind of data, depend on dedicated graphic cards to train due to computational requirements. Computers that can train such models are not widely available for researchers and purchasing capable computers adds a significant cost to the experiments.

It is important to emphasize the need to share and harmonize data. It is crucial to have good quality data, to handle missing data adequately, and to utilize at least few instruments in common and apply them similarly across sites. These strategies will not only facilitate the task of building big datasets but also allow replication of positive findings. Another important point regarding BD is the assessment of rare events, such as suicide. Because of statistical rarity of suicide deaths in the short term, even models with good accuracy would result in a poor risk stratification tool.[95] In this case, resampling strategies should be applied in the training phase.

Multimodality

Working with multimodal data can provide another challenge. There is no standard way of integrating information from multiple sources, such as using both text and image for predicting an outcome. By using features extracted from each modality, the model could possibly use the data, but this kind of approach usually diminishes the potential of information that could be extracted. This happens because feature extraction algorithms most of the time depend on human knowledge, decreasing the potential for data-driven approaches to find hidden patterns.

However, deep learning poses a possible solution for such task. By learning the feature extraction process, deep learning models tend to properly find most of the important information for solving the task. This includes finding the potentially hidden patterns that could be used for clinical practice or leveraging our knowledge on a disorder. To handle multimodal data, the models learn a joint latent space where semantically similar data from any of the modalities presented are close.

This type of analysis does not come without drawbacks. By performing analysis on raw data, deep learning is very susceptible to *domain shift*. By training with almost no preprocessing, the data can follow very different distributions when varying equipment. A good example could be the range of distributions that the same slice of a brain scan can assume based on which MRI protocol and equipment was used. When changing modalities such as MRI or CT, the distribution shifts even more drastically, generally completely invalidating using the model without further training. For that reason, when deploying for clinical practice, the researchers should be careful to guarantee that the scenario in which the model is being applied follows the same distribution from the training data. If this is not maintained, there is no way to assure a good predictive performance.

Non-stationary distribution of the data

Researchers should be careful with non-stationary distributions to use their models. These are cases where the data distribution changes over time, and thus the moment in which the data was collected heavily influences the model's behavior. An example could be a model to predict suicide attempts in patients with BD through their posts in social media. In that case, the relationship between the words and suicide attempts may change in part because the writing style shifts over time. This is one of the most challenging problems in machine learning, as the variables that could impact the model's predictive power are most of the time hidden from our perspective. This also happens very often in companies trying to bring machine learning to production. A real case in healthcare research dates back to 2009 when a model trained for predicting flu epidemics was found by 2011 to consistently overestimate flu prevalence in subsequent years.[96,97]

Heterogeneity and phenomenological diagnosis

Current psychiatric diagnosis is based on clinical judgment of patient's narratives and behavior. Specifically, BD illustrates the dilemma of diagnostic systems solely based on clinical judgment. Clinical observations are subjective in nature, often incomplete, and prone to inconsistencies between evaluators. This scenario hampers the training process in machine learning studies since both outcomes and predictors may be subject to these inconsistencies. As abovementioned, heterogeneity could also complicate this scenario. Also, machine learning, almost by design is suited better for detection of aggregates of multiple small effects. This could be the underlying reality of psychiatric disorders, but it could be also a "diluting effect" of heterogeneity.[98]

Funding agencies and lack of interpretability

Due to their disruptive nature, data-driven approaches are still encountering some acceptance barriers in the healthcare community. Unlike other fields, such as computer science, there is no clear consensus among researchers as to whether such methods are reliable. The main concern is that not fully depending on hypothesis leads to fishing expeditions, and thus the risk of false positive findings. Funding agencies, consequently, are still concerned and conservative in their investment in such approaches. This will probably be overcome as the machine learning protocols for healthcare become more widespread, and the results are more convincing.

Yet another concern about such approaches is the lack of interpretability that the resulting models usually possess. There is no clear way of interpreting complex non-linear models. However, if a phenomenon is presented in a non-linear pattern, the ability to model its function surpasses the need for understanding its behavior. In many clinical applications, it can be expected that the clinician does not need to fully comprehend how the machine is processing information. In that case, the main concern is how effectively the model can predict a specific outcome. Although visualization and interpretation are important, perhaps we should not be too limited by human capabilities of pattern recognition, and instead, be open and interested in how we can improve practice using proven and reliable predictions.

Ethical issues

Big data analytics and machine learning do not come without ethical worries. First, regarding the privacy and anonymity of the data.[99] Hospitals and institutions need to establish clear policies to determine who is granted access to collected information, to avoid sensitive data to be inadequately exposed and analyzed. At the same time, it is our opining that the lack of data sharing is one of the main obstacles, which hinders the full realization of the potential of machine learning. Without large sample sizes and multi-site data, it will not be possible to build reliable machine learning algorithms applicable to heterogeneous psychiatric disorders. Thus, developing ethico-legal framework, which would facilitate safe data sharing is a key and critical component of the machine learning field. The cost of data sharing to participant needs to be carefully weighted against the potential benefits to this approach to the society. In case of brain imaging data, for instance, the benefit of data sharing, which could yield new diagnostic/prognostic tools, markedly outweighs the risk to the participant whose anonymized data are shared.

Second, the impact those predictive models may have for individuals. If we can predict that a patient with BD will have a more pernicious illness course, that would mean he will make more use of health services, and therefore, may be charged more for a health plan. The argument that subjects identified as at risk of suicide, for instance, could suffer psychosocial prejudice is indeed a major concern,[95] however, this issue can be in part handled by fostering medical confidentiality, which is possible provided that these subjects are not at acute risk of suicide. Additionally, if an individual is predicted to develop bipolar disorder or attempts suicide, how this information influences his quality of life? How will it influence his relationships with his peers? It is possible that the stressful burden of knowing may incur in speeding the disorder installment or even lead to another disorder, such as substance abuse. An important question, therefore, is how our patients may cope with such predictions about their future, and weigh harm and benefit of its use. Future guidelines in the field of bipolar disorder may have to address the problem of “potential patients” not only in terms of therapeutic preventive strategies but also in terms of how to handle the harm related to this prediction.

Conclusion

The high morbidity and mortality related to BD provides the impetus for research into more sophisticated computational approaches for risk prediction, individualized treatment, and prognosis. In this manuscript, we summarized how machine learning techniques and big data analysis may help the field by providing predictive models at the individual level. It is important to note that some of the studies included used machine learning techniques but not big datasets. Additionally, some of the most intriguing results derive from small studies that have yet to be independently replicated. The field of machine learning and big data in BD is still in its infancy and replication of the findings is required. However, technology made available by machine learning and big data analytics gives us the unique opportunity to study the “real patient” and all of the inherent complexity.[100] It is also important to mention that in universal health systems, a wealth of untapped and yet available, person-specific information is attached to every single patient and could be used to build diagnostic tools. Currently, the full scope of individual information is under-utilized and the information value of the sequence and timeframe of events is underdeveloped. Until recent times one major constraint for the use of such wealth of information was the lack of means to analyze it in a coherent way with standard statistical techniques. The emerging field of big data and machine learning provides a framework to deal with such broad and complex datasets in real-time to advance our understanding and treatment of BD.

Acknowledgements:

Ives Cavalcante Passos was supported by funding from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and CAPES (Brazilian Government) and FIPE (Hospital de Clínicas de Porto Alegre). Tomas Hajek was supported by funding from the Canadian Institutes of Health Research (103703, 106469 and 142255), Nova Scotia Health Research Foundation, Dalhousie Clinical Research Scholarship, Brain & Behavior Research Foundation (formerly NARSAD) 2007 Young Investigator and 2015 Independent Investigator Awards, the Ministry of Health, Czech Republic (grants number 16-32791A, 16-32696A). Flávio Kapczinski has received grants or research support from Brain & Behavior Research Foundation (formerly NARSAD), the Stanley Medical Research Institute, CNPq, and CAPES (Brazilian Government).

References

1. Merikangas KR, Akiskal HS, Angst J, et al. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Arch Gen Psychiatry*. 2007;64(5):543-552. doi:10.1001/archpsyc.64.5.543.
2. Mathers CD, Iburg KM, Begg S. Adjusting for dependent comorbidity in the calculation of healthy life expectancy. *Popul Health Metr*. 2006;4:4. doi:10.1186/1478-7954-4-4.
3. Nordentoft M, Mortensen PB, Pedersen CB. Absolute risk of suicide after first hospital contact in mental disorder. *Arch Gen Psychiatry*. 2011;68(10):1058-1064. doi:10.1001/archgenpsychiatry.2011.113.
4. Yatham LN, Kennedy SH, Parikh S V, 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;20(2):97-170. doi:10.1111/bdi.12609.
5. Goodwin GM, Anderson I, Arango C, et al. ECNP consensus meeting. Bipolar depression. Nice, March 2007. *Eur Neuropsychopharmacol*. 2008;18(7):535-549. doi:10.1016/j.euroneuro.2008.03.003.
6. Gitlin MJ, Swendsen J, Heller TL, Hammen C. Relapse and impairment in bipolar disorder. *Am J Psychiatry*. 1995;152(11):1635-1640.
7. Kessing LV, Andersen PK, Vinberg M. Risk of recurrence after a single manic or mixed episode - a systematic review and meta-analysis. *Bipolar Disord*. 2018;20(1):9-17. doi:10.1111/bdi.12593.
8. Passos IC, Mwangi B, Vieta E, Berk M, Kapczinski F. Areas of controversy in neuroprogression in bipolar disorder. *Acta Psychiatr Scand*. April 2016. doi:10.1111/acps.12581.
9. da Costa SC, Passos IC, Lowri C, Soares JC, Kapczinski F. Refractory bipolar disorder and neuroprogression. *Prog Neuropsychopharmacol Biol Psychiatry*. 2015;In Press. doi:10.1016/j.pnpbp.2015.09.005.
10. Kessing L V, Andersen PK. Evidence for clinical progression of unipolar and bipolar disorders. *Acta Psychiatr Scand*. 2017;135(1):51-64. doi:10.1111/acps.12667.
11. Lish JD, Dime-Meenan S, Whybrow PC, Price RA, Hirschfeld RM. The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. *J Affect Disord*. 1994;31(4):281-294.
12. Geddes JR, Calabrese JR, Goodwin GM. Lamotrigine for treatment of bipolar depression: independent meta-analysis and meta-regression of individual patient data from five randomised trials. *Br J Psychiatry*. 2009;194(1):4-9. doi:10.1192/bjp.bp.107.048504.
13. Young AH, McElroy SL, Olausson B, Paulsson B. A randomised, placebo-controlled 52-week trial of continued quetiapine treatment in recently depressed patients with

bipolar i and bipolar II disorder. *World J Biol Psychiatry*. 2014;15(2):96-112. doi:10.3109/15622975.2012.665177.

14. Greenhalgh T, Howick J, Maskrey N. Evidence based medicine: a movement in crisis? *Bmj*. 2014;348(jun13 4):g3725-g3725. doi:10.1136/bmj.g3725.
15. Kapczinski NS, Mwangi B, Cassidy RM, et al. Neuroprogression and illness trajectories in bipolar disorder. *Expert Rev Neurother*. 2017;17(3):277-285. doi:10.1080/14737175.2017.1240615.
16. Grof P, Duffy A, Alda M, Hajek T. Lithium response across generations. *Acta Psychiatr Scand*. 2009;120(5):378-385. doi:10.1111/j.1600-0447.2009.01454.x.
17. Garnham J, Munro A, Slaney C, et al. Prophylactic treatment response in bipolar disorder: results of a naturalistic observation study. *J Affect Disord*. 2007;104(1-3):185-190. doi:10.1016/j.jad.2007.03.003.
18. Duffy A, Goodday S, Passos IC, Kapczinski F. Changing the bipolar illness trajectory. *The lancet Psychiatry*. 2017;4(1):11-13. doi:10.1016/S2215-0366(16)30352-2.
19. Hamilton JE, Passos IC, de Azevedo Cardoso T, et al. Predictors of psychiatric readmission among patients with bipolar disorder at an academic safety-net hospital. *Aust New Zeal J Psychiatry*. 2016;50(6):584-593. doi:10.1177/0004867415605171.
20. Savage N. Machine learning: Calculating disease. *Nature*. 2017;550(7676):S115-S117. doi:10.1038/550S115a.
21. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181.
22. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning in psychiatry. *The Lancet Psychiatry*. 2016;3:13-15.
23. Librenza-Garcia D, Kotzian BJ, Yang J, et al. The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neurosci Biobehav Rev*. 2017;80:538-554. doi:10.1016/j.neubiorev.2017.07.004.
24. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*. 2009;339:b2700.
25. Marr B. *Big Data : Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*.
26. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela J-P. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacology*. 2018;43(8):1660-1666. doi:10.1038/s41386-018-0030-z.
27. Kessing LV, Munkholm K, Faurholt-Jepsen M, et al. The Bipolar Illness Onset study: research protocol for the BIO cohort study. *BMJ Open*. 2017;7(6):e015462. doi:10.1136/bmjopen-2016-015462.
28. Munkholm K, Vinberg M, Pedersen BK, Poulsen HE, Ekstrøm CT, Kessing L V. A multisystem composite biomarker as a preliminary diagnostic test in bipolar disorder. *Acta Psychiatr Scand*. 2019;139(3):227-236. doi:10.1111/acps.12983.
29. Mitchell TM (Tom M. *Machine Learning*. McGraw-Hill; 1997.
30. Mwangi B, Tian TS, Soares JC. A Review of Feature Reduction Techniques in Neuroimaging. *Neuroinformatics*. 2013;12(2):229-244. doi:10.1007/s12021-013-9204-3.
31. Bishop CM. *Pattern Recognition and Machine Learning*. Springer; 2006.
32. Passos IC, Mwangi B, Cao B, et al. Identifying a clinical signature of suicidality among patients with mood disorders: a pilot study using a machine learning approach. *J Affect Disord*. 2016;193:109-116. doi:10.1016/j.jad.2015.12.066.
33. Wu M-J, Mwangi B, Bauer IE, et al. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. *Neuroimage*. 2016;In Press. doi:10.1016/j.neuroimage.2016.02.016.
34. Hajek T, Cooke C, Kopecek M, Novak T, Hoschl C, Alda M. Using structural MRI to identify individuals at genetic risk for bipolar disorders: a 2-cohort, machine learning

study. *J Psychiatry Neurosci*. 2015;40(5):316-324.

35. Redlich R, Almeida JJR, Grotegerd D, et al. Brain morphometric biomarkers distinguishing unipolar and bipolar depression. A voxel-based morphometry-pattern classification approach. *JAMA psychiatry*. 2014;71(11):1222-1230. doi:10.1001/jamapsychiatry.2014.1100.
36. Rocha-Rego V, Jogia J, Marquand AF, Mourao-Miranda J, Simmons A, Frangou S. Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach. *Psychol Med*. 2014;44(3):519-532. doi:10.1017/S0033291713001013.
37. Nunes A, Schnack HG, Ching CRK, et al. Using structural MRI to identify bipolar disorders - 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Mol Psychiatry*. August 2018. doi:10.1038/s41380-018-0228-9.
38. Deng F, Wang Y, Huang H, et al. Abnormal segments of right uncinate fasciculus and left anterior thalamic radiation in major and bipolar depression. *Prog Neuropsychopharmacol Biol Psychiatry*. 2018;81:340-349. doi:10.1016/j.pnpbp.2017.09.006.
39. Chuang L-C, Kuo P-H. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. *Sci Rep*. 2017;7(1):39943. doi:10.1038/srep39943.
40. Acikel C, Aydin Son Y, Celik C, Gul H. Evaluation of potential novel variations and their interactions related to bipolar disorders: analysis of genome-wide association study data. *Neuropsychiatr Dis Treat*. 2016;12:2997-3004. doi:10.2147/NDT.S112558.
41. Dmitrzak-Weglarz MP, Pawlak JM, Maciukiewicz M, et al. Clock gene variants differentiate mood disorders. *Mol Biol Rep*. 2015;42(1):277-288. doi:10.1007/s11033-014-3770-9.
42. Laksshman S, Bhat RR, Viswanath V, Li X. DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning. *Hum Mutat*. 2017;38(9):1217-1224. doi:10.1002/humu.23272.
43. Pirooznia M, Seifuddin F, Judy J, et al. Data mining approaches for genome-wide association of mood disorders. *Psychiatr Genet*. 2012;22(2):55-61. doi:10.1097/YPG.0b013e32834dc40d.
44. Struyf J, Dobrin S, Page D. Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. *BMC Genomics*. 2008;9:531. doi:10.1186/1471-2164-9-531.
45. Johannesen JK, O'Donnell BF, Shekhar A, McGrew JH, Hetrick WP. Diagnostic Specificity of Neurophysiological Endophenotypes in Schizophrenia and Bipolar Disorder. *Schizophr Bull*. 2013;39(6):1219-1229. doi:10.1093/schbul/sbs093.
46. Tekin Erguzel T, Tas C, Cebi M. A wrapper-based approach for feature selection and classification of major depressive disorder-bipolar disorders. *Comput Biol Med*. 2015;64:127-137. doi:10.1016/j.compbiomed.2015.06.021.
47. Erguzel TT, Sayar GH, Tarhan N. Artificial intelligence approach to classify unipolar and bipolar depressive disorders. *Neural Comput Appl*. 2016;27(6):1607-1616. doi:10.1007/s00521-015-1959-z.
48. Alimardani F, Cho J-H, Boostani R, Hwang H-J. Classification of Bipolar Disorder and Schizophrenia Using Steady-State Visual Evoked Potential Based Features. *IEEE Access*. 2018;6:40379-40388. doi:10.1109/ACCESS.2018.2854555.
49. Khodayari-Rostamabad A, Reilly JP, Hasey G, Debruin H, Maccrimmon D. Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model. *Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf*. 2010;2010:4006-4009. doi:10.1109/IEMBS.2010.5627998.
50. Lithgow BJ, Moussavi Z, Gurvich C, Kulkarni J, Maller JJ, Fitzgerald PB. Bipolar disorder in the balance. *Eur Arch Psychiatry Clin Neurosci*. August 2018. doi:10.1007/s00406-018-0935-x.

51. Wu M-J, Passos IC, Bauer IE, et al. Individualized identification of euthymic bipolar disorder using the Cambridge Neuropsychological Test Automated Battery (CANTAB) and machine learning. *J Affect Disord*. 2016;192:219-225. doi:10.1016/j.jad.2015.12.053.
52. Besga A, Gonzalez I, Echeburua E, et al. Discrimination between Alzheimer's Disease and Late Onset Bipolar Disorder Using Multivariate Analysis. *Front Aging Neurosci*. 2015;7:231. doi:10.3389/fnagi.2015.00231.
53. Pinto JV, Passos IC, Gomes F, et al. Peripheral biomarker signatures of bipolar disorder and schizophrenia: A machine learning approach. *Schizophr Res*. January 2017. doi:10.1016/j.schres.2017.01.018.
54. Schulz SC, Overgaard S, Bond DJ, Kaldate R. Assessment of Proteomic Measures Across Serious Psychiatric Illness. *Clin Schizophr Relat Psychoses*. 2017;11(2):103-112. doi:10.3371/CSRP.SSSO.071717.
55. Haenisch F, Cooper JD, Reif A, et al. Towards a blood-based diagnostic panel for bipolar disorder. *Brain Behav Immun*. 2016;52:49-57. doi:10.1016/j.bbi.2015.10.001.
56. Munkholm K, Peijs L, Vinberg M, Kessing L V. A composite peripheral blood gene expression measure as a potential diagnostic biomarker in bipolar disorder. *Transl Psychiatry*. 2015;5(8):e614. doi:10.1038/tp.2015.110.
57. Tran T, Kavuluru R. Predicting mental conditions based on &quot;history of present illness&quot; in psychiatric notes with deep neural networks. *J Biomed Inform*. 2017;75S:S138-S148. doi:10.1016/j.jbi.2017.06.010.
58. Yang T-H, Wu C-H, Huang K-Y, Su M-H. Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. *J Ambient Intell Hum Comput*. 2016;8(6):895–906. doi:10.1007/s12652-016-0395-y.
59. Huang K-Y, Wu C-H, Su M-H, Chou C-H. Mood disorder identification using deep bottleneck features of elicited speech. In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE; 2017:1648-1652. doi:10.1109/APSIPA.2017.8282296.
60. Perez Arribas I, Goodwin GM, Geddes JR, Lyons T, Saunders KEA. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Transl Psychiatry*. 2018;8(1):274. doi:10.1038/s41398-018-0334-0.
61. Duffy A, Goodday S, Keown-Stoneman C, Grof P. The Emergent Course of Bipolar Disorder: Observations Over Two Decades From the Canadian High-Risk Offspring Cohort. *Am J Psychiatry*. December 2018:appi.ajp.2018.1. doi:10.1176/appi.ajp.2018.18040461.
62. Mwangi B, Wu M-J, Cao B, et al. Individualized Prediction and Clinical Staging of Bipolar Disorders using Neuroanatomical Biomarkers. *Biol psychiatry Cogn Neurosci neuroimaging*. 2016;1(2):186-194. doi:10.1016/j.bpsc.2016.01.001.
63. Passos IC, Mwangi B, Cao B, et al. Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach. *J Affect Disord*. 2016;193:109-116. doi:10.1016/j.jad.2015.12.066.
64. Niculescu AB, Levey DF, Phalen PL, et al. Understanding and predicting suicidality using a combined genomic and clinical risk assessment approach. *Mol Psychiatry*. August 2015. doi:10.1038/mp.2015.112.
65. Levey DF, Niculescu EM, Le-Niculescu H, et al. Towards understanding and predicting suicidality in women: biomarkers and clinical risk assessment. *Mol Psychiatry*. 2016;21(6):768-785. doi:10.1038/mp.2016.31.
66. de Ávila Berni G, Rabelo-da-Ponte FD, Librenza-Garcia D, et al. Potential use of text classification tools as signatures of suicidal behavior: A proof-of-concept study using Virginia Woolf's personal writings. DeLuca V, ed. *PLoS One*. 2018;13(10):e0204820. doi:10.1371/journal.pone.0204820.
67. Salvini R, da Silva Dias R, Lafer B, Dutra I. A Multi-Relational Model for Depression Relapse in Patients with Bipolar Disorder. *Stud Health Technol Inform*. 2015;216:741-745.

68. Faurholt-Jepsen M, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. *Transl Psychiatry*. 2016;6(7):e856. doi:10.1038/tp.2016.123.
69. 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*. September 2018. doi:10.1038/s41380-018-0250-y.
70. Chekroud AM, Lane CE, Ross DA. Computational Psychiatry: Embracing Uncertainty and Focusing on Individuals, Not Averages. *Biol Psychiatry*. 2017;82(6):e45-e47. doi:10.1016/j.biopsych.2017.07.011.
71. Chekroud AM. Bigger Data, Harder Questions—Opportunities Throughout Mental Health Care. *JAMA Psychiatry*. 2017;74(12):1183. doi:10.1001/jamapsychiatry.2017.3333.
72. Fleck DE, Ernest N, Adler CM, 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;19(4):259-272. doi:10.1111/bdi.12507.
73. Castro VM, Roberson AM, McCoy TH, et al. Stratifying Risk for Renal Insufficiency Among Lithium-Treated Patients: An Electronic Health Record Study. *Neuropsychopharmacology*. August 2015. doi:10.1038/npp.2015.254.
74. Nzeyimana A, Saunders KE, Geddes JR, McSharry PE. Lamotrigine Therapy for Bipolar Depression: Analysis of Self-Reported Patient Data. *JMIR Ment Heal*. 2018;5(4):e63. doi:10.2196/mental.9026.
75. Stern S, Santos R, Marchetto MC, et al. Neurons derived from patients with bipolar disorder divide into intrinsically different sub-populations of neurons, predicting the patients' responsiveness to lithium. *Mol Psychiatry*. 2018;23(6):1453-1465. doi:10.1038/mp.2016.260.
76. Wade BSC, Joshi SH, Njau S, et al. Effect of Electroconvulsive Therapy on Striatal Morphometry in Major Depressive Disorder. *Neuropsychopharmacology*. 2016;41(10):2481-2491. doi:10.1038/npp.2016.48.
77. Duffy A, Grof P, Robertson C, Alda M. The implications of genetics studies of major mood disorders for clinical practice. *J Clin Psychiatry*. 2000;61(9):630-637.
78. Goodday SM, Preisig M, Gholamrezaee M, Grof P, Angst J, Duffy A. The association between self-reported and clinically determined hypomanic symptoms and the onset of major mood disorders. *BJPsych open*. 2017;3(2):71-77. doi:10.1192/bjpo.bp.116.004234.
79. Hafeman DM, Merranko J, Axelson D, et al. Toward the Definition of a Bipolar Prodrome: Dimensional Predictors of Bipolar Spectrum Disorders in At-Risk Youths. *Am J Psychiatry*. 2016;173(7):695-704. doi:10.1176/appi.ajp.2015.15040414.
80. Duffy A, Keown-Stoneman CD, Goodday SM, et al. Daily and weekly mood ratings using a remote capture method in high-risk offspring of bipolar parents: Compliance and symptom monitoring. *Bipolar Disord*. November 2018:bdi.12721. doi:10.1111/bdi.12721.
81. Duffy A, Horrocks J, Doucette S, Keown-Stoneman C, McCloskey S, Grof P. The developmental trajectory of bipolar disorder. *Br J Psychiatry*. 2014;204(2):122-128. doi:10.1192/bjp.bp.113.126706.
82. Jansen K, Cardoso TA, Fries GR, et al. Childhood trauma, family history, and their association with mood disorders in early adulthood. *Acta Psychiatr Scand*. 2016;134(4):281-286. doi:10.1111/acps.12551.
83. Goodday S, Levy A, Flowerdew G, et al. Early exposure to parental bipolar disorder and risk of mood disorder: the Flourish Canadian prospective offspring cohort study. *Early Interv Psychiatry*. 2018;12(2):160-168. doi:10.1111/eip.12291.
84. Doucette S, Horrocks J, Grof P, Keown-Stoneman C, Duffy A. Attachment and temperament profiles among the offspring of a parent with bipolar disorder. *J Affect Disord*. 2013;150(2):522-526. doi:10.1016/j.jad.2013.01.023.
85. Grande I, Berk M, Birmaher B, Vieta E. Bipolar disorder. *Lancet (London, England)*. September 2015. doi:10.1016/S0140-6736(15)00241-X.

86. Birmaher B, Merranko JA, Goldstein TR, et al. A Risk Calculator to Predict the Individual Risk of Conversion From Subthreshold Bipolar Symptoms to Bipolar Disorder I or II in Youth. *J Am Acad Child Adolesc Psychiatry*. 2018;57(10):755-763.e4. doi:10.1016/J.JAAC.2018.05.023.

87. Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF. Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biol psychiatry Cogn Neurosci neuroimaging*. 2016;1(5):433-447. doi:10.1016/j.bpsc.2016.04.002.

88. Bansal R, Staib LH, Laine AF, et al. Anatomical brain images alone can accurately diagnose chronic neuropsychiatric illnesses. Zhan W, ed. *PLoS One*. 2012;7(12):e50698. doi:10.1371/journal.pone.0050698.

89. Hall M-H, Smoller JW, Cook NR, et al. Patterns of deficits in brain function in bipolar disorder and schizophrenia: a cluster analytic study. *Psychiatry Res*. 2012;200(2-3):272-280. doi:10.1016/j.psychres.2012.07.052.

90. Nguyen T, O'Dea B, Larsen M, Phung D, Venkatesh S, Christensen H. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimed Tools Appl*. 2017;76(8):10653-10676. doi:10.1007/s11042-015-3128-x.

91. Wu M-J, Mwangi B, Bauer IE, et al. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. *Neuroimage*. 2017;145(Pt B):254-264. doi:10.1016/j.neuroimage.2016.02.016.

92. Wahlund B, Grahn H, Sääf J, Wetterberg L. Affective disorder subtyped by psychomotor symptoms, monoamine oxidase, melatonin and cortisol: identification of patients with latent bipolar disorder. *Eur Arch Psychiatry Clin Neurosci*. 1998;248(5):215-224.

93. Chekroud AM, Koutsouleris N. The perilous path from publication to practice. *Mol Psychiatry*. 2018;23(1):24-25. doi:10.1038/mp.2017.227.

94. Mwangi B, Ebmeier KP, Matthews K, Steele JD. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain*. 2012;135(Pt 5):1508-1521. doi:10.1093/brain/aws084.

95. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction Models for Suicide Attempts and Deaths. *JAMA Psychiatry*. March 2019. doi:10.1001/jamapsychiatry.2019.0174.

96. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012-1014. doi:10.1038/nature07634.

97. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science*. 2014;343(6176):1203-1205. doi:10.1126/science.1248506.

98. Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M. The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. Reif A, ed. *PLoS One*. 2013;8(10):e76295. doi:10.1371/journal.pone.0076295.

99. Wilson S. Big data held to privacy laws, too. *Nature*. 2015;519(7544):414-414. doi:10.1038/519414a.

100. Lee Y, Ragguett R-M, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *J Affect Disord*. 2018;241:519-532. doi:10.1016/j.jad.2018.08.073.

Table 1. Important terms and their respective definitions

Term	Definition
Double Dipping	Circular analysis of the test data set. The researcher uses the test data twice, overestimating the predictive power
Overfitting	Model too adjusted to the training set. This leads to problems of generalization
Underfitting	Model not well adjusted to the data. This usually implies that a more complex model is required
Feature Selection	Selecting the most important predictors, either by domain experts or in a data-driven manner
Hyperparameter Tuning	Model induction algorithms usually rely on hyperparameters. These are chosen by the user and can be optimized for better results
Kernel	Kernel is a way of computing the dot product of two vectors “x” and “y” in some (possibly very high dimensional) feature space. Kernel methods are a class of algorithms for pattern analysis, whose best-known member is the support vector machine
Internal Validity	Validating results in the same context (data from the same trial, institution, and others)
External Validity	Validating results from different contexts. This shows that the model is able to handle other scenarios (different trials, institutions, and others)
Multimodal Data	Combining data from multiple heterogeneous sources. This includes combining text, audio, video, and others.
Unstructured Data	Data that does not follow a specific organization, such as text and images.
Curse of dimensionality	When multiple features are present, separating data in the multidimensional space becomes easier. Consequently, overfitting data is a common occurrence.
Holdout	Subset of the data that is kept away from the analysis for posterior testing.
Cross-validation	Validation protocol that involves splitting data into *k*-folds to verify if the model is able to generalize from training
Bootstrap	Alternative validation protocol that creates multiple randomized subsets of training data

Table 2. Important points to be considered in machine-learning-based studies.

Methodological feature	Considerations
Representativeness of the sample	Was the study truly representative of the target population heterogeneity or included a selected group of users?
In the case of supervised machine learning studies, the subjects in different groups are comparable based on the study design or analysis.	Did the study control for the most important confounding factors?
Assessment of the outcome	Independent blind assessment, medical record or self-report?
Machine learning approach	Was the machine learning algorithm used to analyze data clearly described and appropriate? Were metrics of performance presented?
Class Imbalance	How did authors address the class imbalance problem?
Test Dataset	Was the test dataset "unseen"?
Feature Selection and Hyperparameter Tuning	Did the study describe both feature selection and hyperparameter tuning?
Missing Data	Did the study describe how authors handled missing data, including if they were inputted or removed?

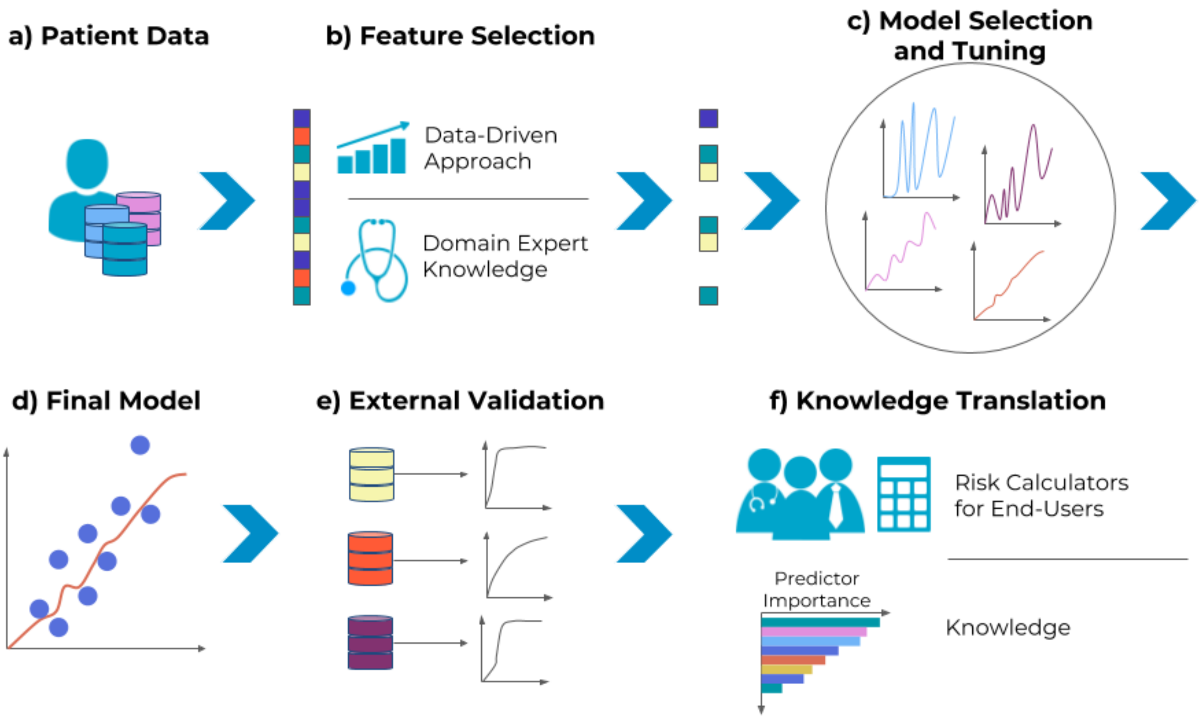

Figure 1. Essential steps to conducting machine learning models. A) The patient data comes from multiple sources and biological levels. B) The most important features should be selected in order to reduce the dimensionality of the problem. This step is done mainly in two different forms. One way is through feature selection algorithms that automatically extract information (data driven). The other is by domain experts that identify which features should be kept through their knowledge on the subject (hypothesis driven). C) Candidate models generated by the induction algorithm. D) Final model chose from the candidate pool by a performance metric, such as the area under the ROC curve or accuracy. E) Model validation with external data, potentially from different institutions to avoid bias. F) Translate the knowledge to generate risk calculators.