Mental Health Commission of Canada
Commission de la santé mentale du Canada

CADTH

CADTH Health Technology Review

Artificial Intelligence and Machine Learning in Mental Health Services: A Literature Review

June 2021

Rapid Review

This literature review was commissioned by The Mental Health Commission of Canada to address the role of AI in mental health services. This report is a companion to an Environmental Scan, which provides more information on the types and trends of AI either emerging or currently in use for the prevention, diagnosis, or treatment of mental health problems and illnesses, research and development initiatives, and the professional groups and organizations involved in the development or use of these technologies in Canada and internationally [Artificial Intelligence and Machine Learning in Mental Health Services: An Environmental Scan. Ottawa: Canadian Agency for Drugs and Technology in Health (CADTH), Mental Health Commission of Canada (MHCC); 2021 June.]

Authors: Charlotte Wells, Shannon Hill, Charlene Argáez

Cite As: Artificial Intelligence and Machine Learning in Mental Health Services: A Literature Review. Ottawa: Canadian Agency for Drugs and Technology in Health (CADTH), Mental Health Commission of Canada (MHCC); 2021 June.

About CADTH: CADTH is an independent, not-for-profit organization responsible for providing Canada's health care decision-makers with objective evidence to help make informed decisions about the optimal use of drugs, medical devices, diagnostics, and procedures in our health care system.

Funding: CADTH receives funding from Canada's federal, provincial, and territorial governments, with the exception of Quebec. For this project, CADTH received funding from the Mental Health Commission of Canada.

Contact **requests@cadth.ca** with inquiries about this notice or legal matters relating to CADTH services.

Mental Health Commission of Canada | Commission de la santé mentale du Canada

CADTH

Table of Contents

Abbreviations

ADHD	attention-deficit/hyperactivity disorder
AI	artificial intelligence
AUC	area under the curve
BPD	bipolar disorder
CAT	computerized adaptive test
DSM	*Diagnostics and Statistics Manual of Mental Disorders*
DSM-IV-TR	*Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition, Text Revision*
DSM-V	*Diagnostics and Statistics Manual of Mental Disorders, Fifth Edition*
DSM-IV	*Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition*
EMA	ecological momentary assessment
EMI	ecological momentary intervention
fMRI	functional magnetic resonance imaging
GAD-7	Generalized Anxiety Disorder 7-item scale
ICD-10	International Statistical Classification of Diseases and Related Health Problems, 10th Revision
K-NHANES	Korea–National Health and Nutrition Examination Survey
LASSO	least absolute shrinkage and selection operator
LDA	linear discriminate analysis
LGBTQ2+	lesbian, gay, bisexual, transgender, queer or questioning, and two-spirited
MA	meta-analysis
MDD	major depressive disorder
MINI	Mini International Neuropsychiatric Interview
ML	machine learning
MRI	magnetic resonance imaging
NHANES	National Health and Nutrition Examination Survey
NLP	natural language processing
NPV	negative predictive value
PHQ-9	Patient Health Questionnaire-[9]
PPD	postpartum depression
PPV	positive predictive value
PSQ	Perceived Stress Questionnaire
PTSD	post-traumatic stress disorder
RCT	randomized controlled trial
RF	random forest
SCZ	schizophrenia
sMRI	structural magnetic resonance imaging
SR	systematic review
SVM	support vector machine

Context and Policy Issues

The burden of mental illness or other mental health problems is high in Canada. One in five people are affected by mental health issues in a given year and one in two people will be affected by mental health issues by the time they are 40 years old.[1,2] Mental health conditions include but are not limited to major depressive disorders (MDD; and other depression-related disorders), anxiety disorders, and schizophrenia (SCZ).[2] Despite specific diagnostic categories for mental illnesses, individuals who have mental health issues often fall on a spectrum of severity that may not always fit into a *Diagnostic and Statistical Manual of Mental Disorders (DSM)* category, or individuals who have mental health issues may have multiple overlapping diagnoses (e.g., anxiety coupled with depression).[3] Barriers to proper care, including stigma, discrimination, accessibility, cost, and lack of flexibility of treatment (such as personalization of treatment or tailoring of treatment to individuals' mental health needs or issues) have been an issue for many Canadians affected by mental illness.[1]

Artificial intelligence (AI) and machine learning (ML) have become major fields of interest for multiple countries, including Canada.[4] AI includes systems that perform tasks such as problem-solving, reasoning, and recognition, and can act autonomously.[5,6] Generally, AI systems are able to perform tasks and functions that require problem-solving and reasoning, and are able to learn from data.[6] ML, a subset of AI, learns from previous data to develop models that can perform functions including classification, regression, clustering, and normative modelling of data.[7] Different types of ML include supervised learning, unsupervised learning, and reinforcement learning.[8] An ML model can essentially learn how to label new information based on previous experiences with similar information, much like the average human individual. For example, an average human can look at a picture and determine what the picture is of, based on seeing other pictures of similar items. ML can work in a similar way. More information on these types of ML are included in a companion report: *Artificial Intelligence and Machine Learning in Mental Health Services: An Environmental Scan*,[9] and a glossary of relevant terms is included in Appendix 1. AI and ML algorithms are being developed to potentially predict, diagnose, or treat mental health issues. AI is able to classify patients based on variables inputted into the model (as in diagnostics), and to predict future prognosis or progression of mental health illnesses at an individual level.[7] In this case, the ML model would work similarly to a psychiatrist or clinician. A clinician who is assessing a patient with a set of symptoms or behaviours may have seen these symptoms before, either in other individuals or in references, and can determine that this likely means an individual has a particular mental health diagnosis. An ML model may be able to do a similar task by taking information inputted about the individual and producing an accurate mental health diagnosis based on previously seen information about other patients. Similarly, an ML model could predict the likely prognosis of a patient's mental health diagnosis based on the prognoses of previously analysed patients with the same or similar mental health diagnoses.

Different applications of ML have currently been developed that are aimed at providing diagnosis and treatment, including computerized adaptive tests (CAT) for diagnosis[10] natural language processing (NLP) in predictive and diagnostic analytics (such as in social media research11)[,12] and conversational agents for therapeutic treatment.[1,13] There are also other applications of AI for mental health, such as companion robots for seniors.[14,15]

The purpose of this report is to evaluate the evidence regarding the populations, primary users, effectiveness, and evidence-based guidelines for the use of AI or ML for the prevention, diagnosis, or treatment of mental health problems or illnesses. It is published as

a companion report to *Artificial Intelligence and Machine Learning in Mental Health Services: An Environmental Scan*,[9] which provides more information on the types and trends of AI or ML either emerging or currently in use for the prevention, diagnosis, or treatment of mental health problems and illnesses; research and development initiatives; and the professional groups and organizations involved in the development or use of these technologies in Canada and internationally.[9]

Research Questions

1. What are the populations for whom artificial intelligence technologies have been applied for the prevention, diagnosis, or treatment of mental health problems or illnesses?
2. Who are the primary users of artificial intelligence technologies applied for the prevention, diagnosis, or treatment of mental health problems or illnesses?
3. What is the main purpose and what are the trends regarding the use of artificial intelligence technologies applied for the prevention, diagnosis, or treatment of mental health problems or illnesses?
4. What is the effectiveness of artificial intelligence or machine learning for the prevention, diagnosis, or treatment of mental health problems or illnesses?
5. What are the evidence-based guidelines regarding the use of artificial intelligence or machine learning for the management of mental health problems and illnesses?

Key Findings

Thirty-four studies were identified that were relevant for this report. Eight studies were systematic reviews (SRs), three were randomized controlled trials (RCTs), and 23 were non-randomized studies. No relevant evidence-based guidelines were identified.

The studies included a variety of populations, including individuals with bipolar disorder, schizophrenia, MDD, postpartum depression, post-traumatic stress disorder, and individuals who have suicidal ideation or have attempted suicide. No specific information on subgroups (such as immigrant, refugee, ethnocultural, or racialized individuals; or First Nations, Métis, Inuit; or lesbian, gay, bisexual, transgender, queer or questioning, and two-spirited [LGBTQ2+]) were found. Two studies focused on young children (ages three to seven), and one study used the National Health and Nutrition Examination Survey (NHANES), which includes children and adults. No effectiveness or accuracy information was found on adolescents or older adults with mental health conditions, as the majority of studies focused on adults over the age of 18 and under the age of 65. Intended users of these AI technologies were primarily clinicians (for diagnosis), but three studies examined models that were intended for use by patients. The primary purpose of the AI or ML models was to differentiate patients who have or do not have mental health conditions or to assist in treatment. Diagnostic accuracy of AI or ML models was generally moderate to high when compared with physician assessment, and AI-based applications for the treatment of patients significantly reduced depression symptoms and increased the use of crisis resources in studies that compared various versions of electronic applications for mental health.

Methods

Literature Search Methods

A limited literature search was conducted by an information specialist on key resources including MEDLINE (via OVID), PsycInfo (via OVID), the Cochrane Library, the University of York Centre for Reviews and Dissemination databases, the websites of Canadian and major international health technology agencies, as well as a focused Internet search. The search strategy was comprised of both controlled vocabulary, such as the National Library of Medicine's MeSH (Medical Subject Headings), and keywords. The main search concepts were artificial intelligence and mental health. No filters were applied to limit retrieval by study type. On December 5, 2019, a focused supplemental search was performed to retrieve articles on AI and loneliness in older people. Where possible, retrieval was limited to the human population. The main search was also limited to English-language documents published between January 1, 2014, and September 5, 2019.

Selection Criteria and Methods

Two reviewers independently screened citations and selected studies. In the first level of screening, titles and abstracts were reviewed and potentially relevant articles were retrieved and assessed for inclusion. Conflicts were resolved through discussion. Conflicts that were not resolved through discussion were resolved by a third reviewer. The final selection of full-text articles was based on the inclusion criteria presented in Table 1.

Table 1: Selection Criteria

Population	Individuals at risk of mental health problems and illnesses (formal diagnosis not required) Individuals with mental health problems and illnesses (formal diagnosis not required) Subgroups of interest: • Life stage (children, adolescents, emerging adults, adults, older adults)[a] • Population (immigrant, refugee, ethnocultural, and racialized; First Nations, Métis, and Inuit; LGBTQ2+; male identifying individuals, female identifying individuals; those who speak foreign and Indigenous languages)
Intervention	Any artificial intelligence, machine learning, or predictive analytics
Comparator	Usual care provided by any health care provider (including peer support); no treatment or waitlist; alternative AI algorithms
Outcomes	Populations in which the technology has been applied (e.g., life stage and population subgroups of interest) Users of the technology (e.g., physicians, patients, or other groups) Purpose of use (e.g., the mental health problem or illness being managed or other purposes) Effectiveness for prevention, diagnosis, or early intervention or treatment of mental health problems or illnesses (e.g., accuracy, health outcomes, and harms or adverse events)
Study Designs	Health technology assessments, systematic reviews, meta-analyses, randomized controlled trials, non-randomized studies, evidence-based guidelines

AI = artificial intelligence; LGBTQ2+ = lesbian, gay, bisexual, transgender, queer or questioning, and two-spirited.

[a] For the purposes of this report, age categories were based on definitions used in selected papers. In the absence of a definition, the following was used: Children, younger than 12 years old; adolescents, 13 to 17 years old; adults, 18 to 65 years old; older adults, older than 65 years old.

Exclusion Criteria

Articles were excluded if they did not meet the selection criteria outlined in Table 1, they were duplicate publications, or were published prior to 2014. Guidelines with unclear methodologies were also excluded. Studies examining an algorithm that was not clearly an AI, ML, or AI-based predictive analytic algorithm were excluded (e.g., static flow charts).

Critical Appraisal of Individual Studies

The included SRs were critically appraised by one reviewer using A MeaSurement Tool to Assess Systematic Reviews (AMSTAR 2),[16] randomized and non-randomized studies were critically appraised using the Downs and Black Checklist,[17] and diagnostic accuracy studies were critically appraised using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2).[18] Summary scores were not calculated for the included studies; rather, the strengths and limitations of each included study were described narratively. Additionally, the critical appraisal in the present review was limited to appraisal of the attributes of the study design and execution and was not a critical appraisal of the AI or ML methods or algorithms themselves. Commentary on the appropriateness of the individual AI or ML algorithms for their intended use is out of the scope of this review.

Summary of Evidence

Quantity of Research Available

A total of 1,025 citations were identified in the literature search. Following the screening of titles and abstracts, 938 citations were excluded and 87 potentially relevant reports from the electronic search were retrieved for full-text review. Two potentially relevant publications were retrieved from the grey literature search for full-text reviews. Of these potentially relevant articles, 55 publications were excluded for various reasons, and 34 publications met the inclusion criteria and were included in this report. These comprised eight SRs, three RCTs, and 23 non-randomized studies. Appendix 2 presents the PRISMA[19] flowchart of the study selection.

Summary of Study Characteristics

Additional details regarding the characteristics of included publications are provided in Appendix 3. Two of eight SRs had broader inclusion criteria than the present review.[13,20] Colombo et al. included ecological momentary assessments (EMA) or ecological momentary interventions (EMI), including non-AI–based EMA and EMIs.[20] The relevant study within this SR examined an EMI that contained AI-based algorithms.[20] Laranjo et al. included any studies examining AI-based conversational agents using unconstrained natural language input, which included patients with a variety of conditions (including autism, asthma, cancer, and sleep apnea).[13] The relevant studies were those that examined mental health conditions.[13] The characteristics and results of the subset of relevant studies are described throughout the remainder of this report.

Study Design

Systematic Reviews

Eight SRs with relevance to this report were identified.[7,13,20-25] The included SRs were published in 2019,[20,21,23,24] 2018,[7,13,22] and 2017.[25] One SR[23] was an update of a previous SR, published in 2017;[26] only the updated SR was retained, as there was complete overlap between the older SR and the updated SR.[23,26] There was some overlap between the SRs (16 studies); Appendix 6 shows the included primary studies in each review, and the degree of overlap between the SRs. The date ranges for the searches were inception of the database to 2018[13,21,22,24] or 2019,[20,23] 2000 to 2017,[7] or 2010 to 2017.[25]

The SRs provided results from one study[20] five studies,[13] 35 studies,[21,24] 26 studies,[22] 66 studies,[7] 48 studies,[25] and 90 studies.[23] The study designs included in the SRs were peer-reviewed studies regardless of design,[21] controlled studies with a Jadad score of more than three,[24] and quasi-experimental, RCT, and cross-sectional designs.[13] One SR had no exclusions for study design.[20] Two SRs did not specify which designs were eligible and did not report the designs of the included studies.[7,23] Four SRs focused solely on diagnostic accuracy studies.[7,21,23,24]

Twenty-six primary clinical studies were identified.[10,27-51] Of these, three were RCTs (one of which was a factorial RCT[41]) regarding treatment or crises prevention,[33,35,41] and 23 were diagnostic accuracy studies.[10,27-32,34,36-40,42-51]

For diagnostic accuracy, all of the 23 diagnostic accuracy primary studies were non-random ized.[10,27-32,34,36-40,42-51] Nineteen studies were cross-sectional studies, 11 were cross-sectional diagnostic studies with case-control based selection,[10,28-30,32,34,37,42,43,46,51] and six had cohort-based selection.[36,38-40,45,49] One study was a longitudinal retest-reliability study,[31] two were retrospective accuracy studies,[47,50] and one was a diagnostic longitudinal prediction study.[48]

Country of Origin

The first authors of the included SRs were based in the US,[21] the UK,[25] Spain,[20] Italy,[24] Australia,[13] China,[7] and Canada.[22,23]

The included primary studies were conducted in the US,[10,27,33-35,39-41,46-48,50] South Korea,[28,29,42,43] China,[30,32,49,51] Canada,[37] the UK,[38] India,[44] Germany,[31,45] and Spain.[36]

Patient Population

Systematic Reviews

The SR included patients or data sets of patients with the following conditions:

- non-suicidal self-injury, suicidal ideation, suicide planning, suicide attempt, or suicide death,[21] including patients with bipolar disorder (BPD), MDD, mood disorders, history of self-injury, SCZ, personality disorders, or suicidal ideation, and also including adolescents, adults, undergraduate students, and older adults[21]
- *DSM*-diagnosed past or current MDD[20]
- chronic or newly diagnosed (with DSM-IV, DSM-IV-TR, DSM-V, or International Statistical Classification of Diseases and Related Health Problems, 10th Revision [ICD-10]) SCZ on any episode number; patients may have been taking antipsychotic medications[24]
- patients with mental health conditions (depression, anxiety, PTSD) and autism spectrum disorder;[13] some mental health populations were not specified

- adults (18 years or older) with diagnostic manual diagnosed bipolar or unipolar depression[22]
- adults (18 years or older) with BPD (types I and II)[23]
- patients with MDD[7]
- patients with common mental health disorders (defined by the UK National Institute for Health Care and Excellence), including depression, PPD, PTSD, anxiety, obsessive compulsive disorder, BPD, seasonal affective disorder, eating disorders, SCZ, attention-deficit/hyperactivity disorder (ADHD), sleep disorder, and suicidality.[25]

Primary Studies

The populations in the included RCTs were users of the Koko application (no mental health diagnosis required),[35] adults with depression or anxiety,[41] and adult students attending university in the US.[33]

- The populations in the diagnostic accuracy studies were:
- trauma survivors with or without PTSD[27,34]
- patients with MDD;[28,29,52] one study limited age to between 18 and 60[32]
- patients with generalized anxiety disorder[52]
- adults (18 to 70 years) with mental health issues, not including SCZ[10]
- patients with psychosomatic inpatients[31]
- patients with SCZ[45,47] (including drug-naive SCZ);[37,51] two studies limited age to between 18 to 40 years[30] and 18 to 65 years[44]
- patients with BPD[45]
- women at risk of PPD[36,48]
- patients with schizoaffective disorders[43]
- children (three to seven years old) without a developmental disorder[39,40]
- participants in the King's Centre For Military Health Research longitudinal cohort[38]
- participants in the South Korean- and US-based NHANES and Korea–National Health and Nutrition Examination Survey (K-NHANES) datasets[42]
- graduate students from the Chinese Academy of Science[49]
- patients with ischemic heart disease.[50]

No specific information on subgroups such as immigrant, refugee, ethnocultural or racialized individuals; or First Nations, Métis, or Inuit; or LGBTQ2+ were found. Two studies focused on young children (ages three to seven),[40] and one study used data from the NHANES,[42] which includes both children and adults. No information was found on adolescents or older adults, as the majority of studies focused on adults over the age of 18 and under the age of 65.

Interventions, Purpose of Intervention, and Comparators

Systematic Reviews

The interventions and comparators for the included SRs were as follows:

- ML techniques for the prediction of suicidal-related events. Suicidal-related events included self-injury, ideation, planning, attempt, and death. All types of ML were eligible, but only studies regarding supervised ML techniques (e.g., regularized regression, decision trees, random forest [RF] and black box methods [i.e., accessible input and outputs, but no knowledge of internal workings of the algorithm]) were identified in the search.[21]
- Smartphone-based and handhold technology-based EMAs and EMIs. EMAs and EMIs collect data in real time (usually from sensors or smartphones) and therefore show data within the "natural" environment (in comparison to a laboratory or clinical environment). EMIs provide extended treatment outside of clinical settings.[20] One included EMI (Mobylize!) was an ML-based EMI that predicted the mood state of the user based on attributes such as previous moods, social context, and activities, and sent tailored feedback to the patient.[20] Further information on Mobylize! is provided in Appendix 3.
- ML techniques for the differentiation of SCZ patients from healthy control groups using neuroimaging data (functional magnetic resonance imaging [fMRI] or structural magnetic resonance imaging [sMRI]). Some ML techniques included in the primary studies were ridge, least absolute shrinkage and selection operator (LASSO), elastic net and L0 Norm regularized logistic regression, support vector classifier, regularized discriminant analysis, RF and a Gaussian process classifier, and recursive feature elimination.[24]
- Conversational agents using unrestrained natural language inputs, such as chatbots, embodied conversational agents, and smart conversational interfaces (e.g., Siri, Alexa). The type of ML was not specified.[13]
- Supervised ML techniques or unified modelling language that use predictors (e.g., neuroimaging, phenomenological indicators, genetic indicators, or combined indicators) to predict treatment response (treatments were evidence-based treatments for depression). Classification systems were binary, non-binary, multivariate, and hierarchical clustering algorithms.[22] ML included types of support vector machines (SVM), logistic regression, and artificial neural networks.[22]
- Various ML techniques. The techniques were not specified in the inclusion criteria, but included primary studies used SVM, Gaussian process classifiers, principal component analysis, artificial neural networks, posterior probability model selection, recursive feature elimination, multivariate logistic regressions, Akaike's information criterion, RF, naive Bayes, multifactor dimensionality reduction, decision trees, ensemble of voters, LASSO, and classification and regression trees.
- ML methods for the classification of MDD patients using magnetic resonance imaging (MRI) data (decision trees, SVM, Gaussian process classifiers, LASSO, linear discriminant analysis [LDA]).[7]
- ML techniques that used social media to classify or predict mental health problems. These included LDA, SVM, linguistics inquiry and word count, naive Bayes, artificial neural networks, and LASSO.[25]

Primary Studies

The interventions and comparators for the included RCTs were as follows:

- The Koko application. Koko is a psychoeducation intervention that is designed to reduce perceived barriers to crisis resources. The comparator was a different version of the application.[35]
- The IntelliCare platform. IntelliCare is comprised of 12 applications for a mobile phone designed around psychological treatments, with ML used to tailor treatment options. The comparator was different versions of the application, with some versions containing weekly reminders and some containing specific coaching.[41]
- Tess. Tess is an AI chatbot that provides support through various methods (cognitive behavioural therapy, mindfulness-based therapy, emotionally focused therapy, acceptance and commitment therapy, motivational interviewing, self-compassion therapy, and interpersonal psychotherapy). This was compared with electronic education alone.[33]

The diagnostic accuracy studies included various types of ML algorithms: SVM (either with linear, Gaussian, or multi kernels),[27-29,32,36,38-40,43-45,48,51] neuro-fuzzy networks,[29] LDA,[29] logistic regression,[29,32,36,40,48] Bayesian networks,[29 RF,30,32,38,40,45,46,48] ensemble ML models,[37] artificial and convolutional neural networks,[36,38,43] deep learning,[42] naive Bayes,[36,48] XGBoost,48 k-nearest neighbours,[39] NLP,[34,50] and bagging.[38] Two studies examined CATs (adaptive tests that can use a variety of ML algorithms).[10,31] One study had developed a mobile application (eDPP Predictor) based on the ML algorithm, intended to predict risk of PPD in mothers who had recently given birth.[36]

The reference standards for the diagnostic accuracy studies were physician or psychiatrist diagnosis or assessment,[28-30,34,39,40,43,44,46,51] clinician-administered PTSD scales,[27] ICD-10 criteria or codes,[32,48] Patient Health Questionnaire-9 (PHQ-9) criteria,[10,31,42,49] Generalized Anxiety Disorder 7-item scale (GAD-7) criteria,[10,31,49] the Mini International Neuropsychiatric Interview (MINI) Plus, [37] PTSD civilian checklist,[38] "established diagnoses,"[45] Centers for Medicare & Medicaid Services Hierarchical Condition Categories,[47] Spanish Edinburgh Postnatal Depression Scale test version and diagnostic interview for genetic studies,[36] and manual review of clinician notes.[50]

Outcomes

Systematic Reviews

Eight SRs reported on diagnostic accuracy outcomes.[7,13,20-25] The diagnostic accuracy outcomes were accuracy,[7,13,20-25] sensitivity,[7,21,23,24] specificity,[7,21,23,24] area under the curve (AUC),[21-24 R2,21] receiver operating characteristic,[22] negative predictive value (NPV),[21] positive predictive value (PPV),[21] recall value (i.e., sensitivity),[21] precision error rate,[24] false-positive rate,[23] and false-negative rate.[23]

Three SRs also had other mental health related outcomes.[13,20,22] Mental health outcomes were depression,[13,20] anxiety,[13] and PTSD symptoms;[13] meditation frequency;[13] mood symptom severity;[22] occupational or psychosocial functioning;[22] depression-related hospital admission frequency or duration;[22] and suicidal ideation.[22]

One SR had an outcome that was prediction of mental illness.[25] One SR had an outcome that was acceptability of the intervention.[13 25]

Primary Studies

Outcomes in the RCTs were depressive symptoms measured via the PHQ-[9,33,41] anxiety measured via GAD-7,[33,41] affect measured through the Positive and Negative Affect Schedule,[33] use of crisis resources,[35] satisfaction with service or acceptability,[35] engagement with application,[33,41] and accuracy of assessment.[35]

Outcomes in the diagnostic accuracy studies were accuracy,[27-30,32,34,36,37,39,40,43-46,51,52] sensitivity (or recall),[10,27-30,32,34,36-40,43-45,48-50,52] specificity,[10,27-30,34,36-40,43-45,48,52] NPV,[28-30,34,43] PPV (or precision),[28-32,34,43,49,50] AUC,[36,39,40,42,43,45,48] receiver operating characteristic,[40] F-measure,[49,50] R2,[47] cost accuracy,[47] retest-reliability,[31] true positives,[37] true negatives,[37] false-positives,[37] and false-negatives.[37]

Summary of Critical Appraisal

Additional details regarding the strengths and limitations of the included publications are provided in Appendix 4.

Systematic Reviews

The included SRs ranged in quality. All included SRs performed a comprehensive literature search using two or more databases and provided keywords for the search strategy.[13,20-24 7,25] However, only two SRs searched grey literature,[13,22] and six searched reference lists of included studies.[7,13,22-25]

All included SRs also described the populations eligible for inclusion and the interventions eligible for inclusion,[7,13,20-25] and three of the eight SRs included descriptions of what outcomes were eligible for inclusion in the SR.[21-23] The eligible comparators for the SR and comparators within the primary studies were not reported for many of the SRs, which made it unclear what the active ML or AI intervention was being compared with. This was especially prevalent in SRs of diagnostic accuracy studies, in which calculation of the outcome measures of sensitivity, specificity, or accuracy requires a reference standard (i.e., a "gold standard" test that determines the presence or absence of the condition). The lack of detail on the reference standards made interpretation of accuracy measures difficult both within and across studies.

De Filippis et al. and Gao et al. were explicit in the characteristics of participants who were eligible for inclusion (for both those with and those without a formal mental illness diagnosis).[7,24] Some SRs did not clearly report patient characteristics; therefore, it was not possible to determine whether accuracy outcome measures were calculated from patient data sets with only healthy controls, or from more diverse datasets (e.g., patients with multiple mental disorders). This limits the interpretability of the accuracy of the outcome measures. For example, in the context of BPD, clearly reported patient characteristics help in determining if the included intervention only classifies patients as bipolar or healthy (dichotomously), or also differentiates between bipolar and unipolar depression, SCZ, or other mental health diagnoses to assist in differential diagnoses.

One SR included a meta-analysis (MA).[22] It was unknown whether the MA was appropriate — the MA had extremely high and significant heterogeneity (92%).[22] There was no justification for the MA or why these studies were appropriate to be combined. Many of the combined studies had different predictors, treatments, and interventions, and it may not have been appropriate to pool their data to create a summary statistic.[22]

Finally, within the SRs, many included studies used the same data sets to train and validate the algorithms as they did to test the classification accuracy. Algorithms that are tested and created on the same data sets should be corrected for overfitting.[22,23] It was unclear if the included primary diagnostic accuracy studies within the SRs corrected the diagnostic accuracy estimates for overfitting, which may have overinflated the reported accuracy outcomes.[22,23] However, Lee et al. and Passos et al.[22,23] were the only SRs to mention this as a limitation of their included studies; therefore, it is not possible to evaluate whether this occurred in all included SRs. [22,23]

Primary Studies

The included RCTs all included clear objectives, descriptions of interventions and outcomes, and findings.[33,35,41] Mohr et al. and Fulmer et al. included power calculations and enrolled appropriate sample sizes.[33,41] All included RCTs used appropriate randomization techniques, and none of the studies included blinding, but this lack of blinding was unlikely to affect the results of the studies.[33,35,41] It was unlikely that any of the participants came in contact with one another and it was unlikely that the participants were aware of alternative versions of the applications. All of the data were gathered through the applications and were not gathered in a way that could have introduced measurement bias.

One limitation of the RCTs was the external validity of some results. For example, in Mohr et al., participants were only recruited if they had access to an Android-based phone (due to application compatibility). This limits the conclusions that can be made about a broader population, as a large population group (e.g., iPhone users) were excluded.[41] Similarly, one RCT relied on recruitment of current users of the Koko application, which limits the generalizability of results to individuals who may not have been aware of the application, or who were aware of the application but did not participate in its use.[35] Additionally, attrition bias may have occurred in one RCT; in Jaroszewski et al., over 50% of participants did not respond to follow-up requests. The characteristics of those who did not respond to the follow-up request were not reported, nor were the reasons for loss to follow-up, and therefore it was unknown if these individuals differed systematically from those who did respond. Loss to follow-up in the other two RCTs was low.[33,41]

The included RCTs did not take confounding into account.[33,35,41] For example, in Fulmer et al., the control group had higher anxiety at baseline when compared with the intervention group, which was not controlled for in the analysis.[33] One study acknowledged that individuals in the control group experienced an increase in anxiety that may have been due to confounding variables that could have been adjusted (these confounding variables were not reported).[33] This may have been due to the control intervention directly increasing awareness of a participant's anxiety, leading to higher reported anxiety after the follow-up period.[33] This may lead to bias in the results of the study, and perhaps overestimation of the effect of the intervention.

The majority of the included diagnostic accuracy studies employed a case-control based selection method, in which both cases (i.e., individuals who are diagnosed with a mental health disorder) and controls (i.e., those not diagnosed with a mental health disorder) were recruited. Diagnostic studies that use this methodology may overestimate the accuracy of the index tests (i.e., in this case the ML methods) through a type of sampling bias termed spectrum bias. Spectrum bias refers to a bias that occurs in diagnostic accuracy where the performance of a diagnostic test will vary based on which populations are being tested (i.e., varying prevalence of cases and characteristics of population). Case-control studies can

also overinflate the prevalence of some mental health conditions; for example, in a real-world setting the prevalence of SCZ is approximately 1%,[53] but in case-control–designed studies, the number of SCZ cases can be over 50% of the population being studied. This may not directly affect some diagnostic accuracy outcomes (such as sensitivity and specificity) when applied to a real-world setting with a different prevalence of mental health illness, but other diagnostic accuracy outcomes (such as PPV and NPV) may be affected. Additionally, in some studies, individuals who had mental health conditions were on medications that may have affected variables used in the index test (e.g., antipsychotic drugs, which may affect the brain and therefore MRIs). This may have produced differential results in the patients with mental illnesses that were a result of the medication, and not the presence of mental illness. This confounder could have affected the accuracy of the ML algorithms, as the algorithms may have analyzed these confounders to find differences between the groups. Two studies specifically used drug-naive schizophrenic patients to avoid this bias.[37,51]

All of the diagnostic accuracy studies used appropriate reference standards for the comparators; the most common reference standard was an evaluation by a psychiatrist or physician and diagnosis using the *DSM-IV* or *DSM-V* criteria. Additionally, the researchers were not blinded to the results of the reference standard prior to performing the index tests, but as the index tests were automated algorithms, this was unlikely to affect the accuracy results. However, the timing of the tests relative to one another was not clear in many studies. Therefore, it was unknown if other factors occurred during this time, such as the individuals' condition worsening (for example, if the reference standard was performed significantly earlier than the collection of variables, this could lead to more easily identifiable mental health conditions by the ML method due to the worsening of the condition).

Summary of Findings

Appendix 5 presents a table of the main study findings and the authors' conclusions.

What are the populations for whom artificial intelligence technologies have been applied for the prevention, diagnosis, or treatment of mental health problems or illnesses?

The populations identified in the studies included in this report varied in age, mental health condition, and condition severity. There were no identified studies examining the prevention of mental health issues using AI technologies.

In diagnosis of mental health problems, the populations examined in the identified SRs were patients who had suicidal ideation, had planned suicide, had self-injured, had attempted suicide or who had died by suicide,[21] patients with *DSM*-diagnosed MDD,[7,20] patients with SCZ,[24] adults with BPD[23] patients with all mental health conditions (which included PTSD, depression, and anxiety),[13,25] and those with bipolar or unipolar depression.22 The populations in the primary studies focusing on diagnosis were patients with PTSD[27,34] MDD[28,29,32,52] general anxiety disorder[52] SCZ[30,44,45,47] (including drug-naive SCZ and schizoaffective disorders)[37,43,51] and BPD.[45]

In the treatment of mental health illnesses, the populations examined in the identified SRs were bipolar or unipolar depression,[22] and patients with mental health issues (which included PTSD, depression, and anxiety).[13] In the three RCTs that focused on treatment, the populations were patients who were using the Koko application (no mental health diagnosis required),[35] adults with depression or anxiety[41] and adult students attending university in the US.33

More specific details on the populations included in the studies can be found in Appendix 3.

Who are the primary users of artificial intelligence technologies applied for the prevention, diagnosis, or treatment of mental health problems or illnesses?

Two of the SRs reported the intended users of the AI technologies. De Filippis et al.[24] examined neuroimaging and diagnoses, and specified the intended users to be individuals involved in diagnostics. Laranjo et al.[13] examined conversational agents and specified the intended users to be consumers, caregivers, or health care professionals. The remaining SRs did not specify their intended users; however; it can be assumed that the likely users of the included algorithms are intended to be individuals who would use the diagnostic information that the algorithm provides, namely clinicians who used the algorithm to diagnose patients or to predict the prognosis of a patient.

In the primary studies, Wang et al. specified intended users to be health organizations or those in planning of treatment, as it attempted to predict potential high-cost patients with SCZ who may therefore need extra treatment or care.[47] Jimenez-Serrano et al. specified intended users to be clinicians or mothers who had just given birth, for algorithms that may predict PPD incorporated into a mobile application (eDPP Predictor).[36] The remaining primary diagnostic accuracy studies did not specify the intended users, however; it can be assumed that the likely users are intended to be individuals who would use the diagnostic information that the algorithm provides, similar to the SRs.[47]

The three included RCTs looked at applications of AI through a mobile device, and were intended for use by the patients themselves as a treatment option or as a resource.[33,35,41]

What is the main purpose of artificial intelligence technologies applied for the prevention, diagnosis, or treatment of mental health problems or illnesses?

The main purposes of the AI technologies applied in the SRs were as follows:

- predicting suicide-related events[21]
- predicting the mood state of the patient as part of an EMI and sending tailored feedback to the patient[20]
- differentiating between patients with SCZ and patients without SCZ (or other mental illness) through neuroimaging (fMRI or sMRI)[24]
- using big data to differentiate patients with BPD from patients without BPD (or other mental illness)[23]
- classifying patients with MDD compared with patients without MDD (or other mental illness) using neuroimaging data (MRI)[7]
- using ML in conversational agents with unconstrained natural language input to interact with patients with mental health conditions[13]
- using neuroimaging, phenomenological, genetic, or combined predictors to predict treatment response in adults with bipolar or unipolar depression;[22] the goal was to determine which patients were likely to be a treatment "responder" versus a treatment " non-responder;" the treatments were evidence-based and guideline-concordant treatments for depression[22]
- using social media to predict or classify an individual's risk of mental health problems or status of mental health diagnoses.[25]

The nature of ML interventions in the RCTs were as follows:

- A risk assessment platform that attempted to reduce perceived barriers to the use of crisis resources through a first-person persona (KokoBot). The ML algorithm assessed the semantic content of posts to classify a patient as "in crisis" or not "in crisis" and reacted accordingly (immediate messaging from the application, and series of assessments to determine the reason for crisis, and to present crisis resources).[35]
- An ML mobile application that collected patient data and adapted both intervention content (lessons and tools delivered via the application) and motivational messaging to the patient and their progress.[41]
- A psychological chatbot that provided mental health support in a conversational format using various styles and types of therapy.[33]

The purpose of the ML algorithms in the diagnostic studies was to properly diagnose patients either with or without a mental health condition, or to attempt to differentiate mental health conditions within one population (i.e., assist with differential diagnoses in a population group in which more than one mental health condition is present, which may reflect real-world use).[10,27-32,34,36-40,42-46,48-51] One study used ML to attempt to predict potential high-cost patients (this study reported the patients in the top 10% and top 20% of average per member per month total costs) with SCZ over a one-year period.[47]

What is the effectiveness of artificial intelligence or machine learning for the prevention, diagnosis, or treatment of mental health problems or illnesses?

More detailed results regarding the effectiveness of AI or ML are provided in Appendix 5.

Suicidality — Diagnosis and Referrals

One SR authored by Burke et al.[21] reported that in suicide prevention, AUCs ranged from 0.71 to 0.89 when predicting death by suicide using regression trees, elastic net regression, and LASSO in service members who had a baseline mental health visit. For suicide attempts, sensitivity ranged from 0.54 to 0.87 and specificity ranged from 0.80 to 0.86. A longitudinal study found that their model's performance increased when closer to the time of the suicide attempt (720 days to seven days prior to attempt).[21] AUC in one identified study for suicidal planning was 0.89. Ten studies examined suicidal ideation, finding sensitivities ranging from 0.47 to 0.[88,] and specificities ranging from 0.57 to 0.94.[21]

In one SR, smartphone-based conversational agents inconsistently responded to the phrase "I want to commit suicide," but some agents appropriately referred the individual to a suicide hotline.[13]

Schizophrenia and Schizoaffective Disorders

Diagnosis

In the SR by de Filippis[24] the included primary studies that examined diagnosis of SCZ reported accuracies ranging from 63.9% to 88.4% for sMRI (using SVM) and 41% to 99.3% for fMRI (using multivariate pattern analysis and extreme learning machines, respectively). The comparator for these accuracies was reported for the most accurate fMRI outcome (99.3%) as "more known ML methods" (P. 1623). Across all primary studies include in de Filippis, the most reported accuracy values for all ML methods examined (detailed in the Interventions, Purpose of Intervention, and Comparators section) were between 75% and 90%, with higher overall accuracies in the fMRI studies.[24]

Five primary studies evaluated the use of ML in patients with SCZ, finding accuracies of 76% compared with structured clinical interviews[30] 86.9% compared with the MINI Plus[37] 64 to 65% compared with structured clinical interviews[44] and 60.48% to 84.29% compared with clinician assessment .[51]

In one primary study, 3-D convolutional autoencoder-convolutional neural networks were used to differentiate schizoaffective disorders from patients with no mental health illnesses with an accuracy of 84.43% when compared with other 3-D convolutional neural network models.[43]

Bipolar Disorder

Diagnosis

In the SR by Passos et al.[23] the primary studies had sample data sets of patients with and without BPD, and classification accuracy using sMRI and diffusion tension imaging ranged from 57% to 100% (no reference standard reported). Classification accuracy using fMRI data as the input ranged from 61.7% to 93.1% with an SVM model. Classification accuracy using genetic analysis data as the input ranged from 53.6% to 73.4% with RFs. When predicting clinical outcomes of BPD, the accuracy of prediction was 85% for depression relapse, 61% to 74% for mood changes, 64.7% to 78.8% for suicide, and 84.9% for hyper-reactivity.[23]

In classification models examining MDD, accuracy ranged from 54.8% to 99%, sensitivity ranged from 71% to 100% (where reported), and specificity ranged from 85% to 86% (where reported).[7]

In the MA performed in Lee et al., pooled estimates of accuracy for mood disorders differed when the models were informed by a single data type versus multiple data types.[22] Integration of multiple data types were the most accurate overall (pooled estimation of classification accuracy = 0.[93,] 95% confidence interval, 0.86 to 0.97).[22]

Major Depressive Disorder and Mood Disorders

Diagnosis

In primary studies, in patients with or without MDD, ML models that used the variable of heart rate variability yielded an accuracy of 74.4% when[28] compared with board-certified psychiatrist diagnosis;[28] ML models that used entropy features of heart rate variability yielded an accuracy of 70% when[29] compared with senior psychiatrist evaluation;[29] and ML models that used electroencephalogram measurements yielded an accuracy or 76.19% to 79.63% when[32] compared with ICD-10 criteria.[32]

In CAT testing, the Computerized Adaptive Test for Major Depressive Disorder was able to identify individuals with MDD with 0.96 sensitivity and 0.64 specificity.[10] A CAT assessment in Devine et al. was able to achieve similar retest-reliability to conventional depression, stress, and anxiety instruments (PHQ-[9,] GAD-[7,] and Perceived Stress Questionnaire [PSQ]).[31]

Deep learning models estimated depression in South Korean and US-based NHANES and K-NHANES datasets with "relatively high accuracy" compared with the PHQ-9.[42]

In ML models analyzing gait patterns using a Microsoft Kinect recording, predictive accuracies for anxiety and depression ranged between -0.07 and 0.51 compared with GAD-7 scores and -0.16 and 0.51 compared with PSQ scores.[49]

In patients with ischemic heart disease, analysis of discharge summaries to determine depression status using NLP achieved F-measures of 89.6% in high confidence cases (i.e., when depression

diagnosis terms were included in discharge notes), and achieved F-measures of 70.6% in intermediate confidence cases (i.e., when a combination of antidepressant treatments, psychiatry consultations, or symptoms of depression were included in discharge summary).[50] The NLP model was compared with coded diagnoses and manual human review.[50]

Treatment

One study was identified in Colombo et al. that examined EMIs. The study involved Mobylize!, a ML EMI that was shown to significantly reduce depressive symptoms in a sample size of seven patients relative to the MINI (not reported in the SR, comparator information from original publication).[20,54] More information on Mobylize! is located in Appendix 3.

In the SR by Laranjo et al., conversational agents significantly reduced depression symptoms (P = 0.04), but did not reduce anxiety or affect when compared with psychotherapy support and education.[13]

Individuals who used the all versions of the IntelliCare platform showed reductions in depression symptoms and anxiety symptoms (within-group comparison, P < 0.001).[41] Inventions with coaching added to the application were significantly more effective in reducing anxiety symptoms (P = 0.03, measured using the GAD-7) when compared with no coaching, but coaching was not significantly more effective in depression symptoms (P = 0.06, measured using the PSQ).[41]

Groups that used Tess over two or four weeks had significant reductions in their depression and anxiety symptoms compared with electronic education alone, and had 86% satisfaction with Tess compared with 60% satisfaction with electronic learning.[33]

Prediction

In social media research, sentiment analysis yielded 80% accuracy when predicting depression, and tweets predicted future behavioural changes in women who had just given birth with an accuracy of 71%.[25]

Post-Traumatic Stress Disorder

Diagnosis

In patients with or without PTSD, "moderate-to-high" accuracies were found when differentiating between individuals with PTSD, trauma survivors without PTSD, and individuals with no trauma[27,34] This was compared with the Clinician-Administered PTSD Scale.

In a UK military cohort, ML methods were able to identify "probable PTSD" with accuracies ranging from 89% to 97% compared with the PTSD civilian checklist.[38]

Postpartum Depression

ML model performance for patients at risk of PPD ranged from AUCs of 0.69 to 0.79 using electronic health records.[48] Naive Bayes models performed the best in predicting PPD post-childbirth, with G values of 0.73, sensitivities of 0.72, and specificities of 0.73.[36] This model was integrated into a clinical depression support system within an Android mobile application.[36]

Other Mental Health Issues or Combined Groups

The Koko application was 93% accurate in determining individuals in crisis. Participants in crisis who were assigned to the intervention group were more likely to access available crisis resources (relative risk = 1.23).[35]

Classification accuracies (AUCs) in differentiating SCZ and BPD from patients with no mental illnesses ranged from 58% to 90%[45] The accuracy (AUC) using voxel-based morphometry RF in patients with SCZ versus patients with no mental illnesses compared to "established diagnoses" ranged from 0.58 to 0.82. The accuracy (AUC) using voxel-based morphometry RF in patients with BPD versus patients with no mental illnesses compared with "established diagnoses" was 0.63 in adult patients.

Mean accuracy (using an RF model) in differentiating a combined group of patients with schizoaffective disorders, BPD, and MDD from patients with no mental health illnesses was 91.9% compared with a diagnostic interview for genetic studies and clinician assessment. The mean accuracy (using an RF model) in differentiating SCZ from other psychiatric disorders (BPD and MDD combined) was 77.8% compared with a diagnostic interview for genetic studies and clinician assessment.[46]

In children without a developmental disorder, ML methods used to analyze speech patterns during the three-minute speech task to diagnose internalizing disorders (Appendix 3) produced classification accuracies ranging from 57% to 80% on "high quality" data (i.e., "moderate to very strong representation of speech content and frequency" [P.3]) compared with multimodal assessments.[40] In children, SVM with linear kernel, decision trees, and k-nearest neighbours models using data from wearable sensors during a fear-induction exercise produced classification accuracies in diagnosis of internalizing disorders between 58% and 69% compared with multimodal assessments.[39]

What are the evidence-based guidelines regarding the use of artificial intelligence or machine learning for the management of mental health problems and illnesses?

No evidence-based guidelines were identified regarding the use of AI or ML for the management of mental health problems and illnesses; therefore, no summary can be provided.

Limitations

The included studies had several limitations, including the lack of separate testing data sets for the AI algorithms. Also, many of the included studies used cross validation but did not provide a separate "unknown" testing data set (a hold-out set) on which to test the models. This may lead to an overestimation of the accuracy of the model.

Additionally, many studies had limited sample sizes that may also lead to overfitting of the data. ML algorithms often require large amounts of data in order to be generalizable to other unseen data sets and therefore to be reliable.[9]

Although there were several studies examining diagnostic accuracy using ML, there were comparatively fewer studies examining applications of ML for prevention or treatment. This may reflect the relatively fewer algorithms available for the treatment of mental health conditions, or the relatively fewer research efforts for this field.[9] Laranjo et al. noted that treatment options such as conversational agents are still in their "infancy," which is reflected in the types of studies identified in their SR (quasi-experimental) and the primary studies' recent publication dates (most published after 2010).[13] There were limited numbers of studies examining a particular algorithm type, as the studies used different ML methods and different variables in the models.

A variety of mental health conditions were represented in the literature; however, mental health problems can vary in severity and can often co-occur in individuals. Evidence regarding

the use of AI technologies in the diagnosis or treatment of mild mental health conditions (such as those that do not meet *DSM-IV* or *DSM-V* criteria), or individuals with co-occurring disorders (such as anxiety and depression) was limited.

Conclusions and Implications for Decision- or Policy-Making

Thirty-four studies were identified that addressed the research questions for this report (eight SRs, three randomized controlled trials, and 23 non-randomized studies). No relevant evidence-based guidelines were identified.

Populations for whom AI technologies have been applied include patients with suicidality, PTSD, MDD, SCZ, PPD, BPD, and anxiety. Intended users of the models were primarily clinicians (for diagnosis), but three primary studies examined the effectiveness of models that were intended for use by patients.

The main purpose of the identified AI technologies were for the diagnosis of mental health conditions. Other purposes included assessment of risk or prediction of mental health illness, prediction of individuals with future high costs to the health care system, and treatment using mobile applications. AI and ML algorithms had moderate-to-high accuracy when differentiating patients with versus without mental health conditions or from other mental health conditions using a variety of different variables, including social media posts, neuroimaging, genetics, electroencephalogram, blood biomarkers, and other variables. Three studies examining AI-based treatment options such as conversational agents and mobile phone applications were identified, reporting that the algorithm-based treatments increased the use of crisis resources, reduced depressive symptoms, and reduced anxiety symptoms.

Limitations of the evidence included the variable quality in SRs; an MA conducted using studies that had different interventions, variables, and comparators (resulting in high heterogeneity); the potential overestimation of accuracy results due to case-control based selection methods; and concerns regarding external validity of the studies. Sample sizes ranged from 60 to 39,450 patients and there was a lack of studies examining mild or transient mental health conditions.

No specific subgroups were identified in the literature, such as immigrant, refugee, ethnocultural, or racialized individuals; or First Nations, Métis, or Inuit; or LGBTQ2+. Two studies focused on young children (ages three to seven), and one study used the NHANES, which includes children and adults. Further research using ML algorithms within a "real-world" scenario with larger sample sizes, with a greater variety of potential mental health conditions, and with cohort-based selection methods may help to reduce uncertainty in the true accuracy of the algorithms. For studies examining treatment effectiveness using AI treatments for mental health, further research using methods of randomization, specific control groups (including usual care, such as psychotherapy), and specific mental health outcomes (such as depressive symptoms) may also reduce uncertainty regarding the role of AI-based treatments in the paradigm of mental health treatment. Moreover, more opportunities to test the algorithms on new data sets will assist in determining the generalizability of algorithms to different populations.

References

1. Ackerman ML, Virani T, Billings B. Digital mental health - innovations in consumer driven care. *Nursing Leadersh.* 2017;30(3):63-72.
2. Canadian Mental Health Association. Fast facts about mental illness. 2019; https://cmha.ca/fast-facts-about-mental-illness. Accessed 2019 Dec 11.
3. Kassraian-Fard P, Matthis C, Balsters JH, Maathuis MH, Wenderoth N. Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Front Psyc.* 2016;7.
4. Morch C, Gupta A, Mishara B. Canada protocol: an ethical checklist for the use of artificial intelligence in suicide prevention and mental health. Montreal (QC): Montreal Ethics Institute; 2018: http://canadaprotocol.com/. Accessed 2019 Dec 13.
5. Reavie V. Do you know the difference between data analytics and AI machine learning? *Forbes*. 2018 Aug; https://www.forbes.com/sites/forbesagencycouncil/2018/08/01/do-you-know-the-difference-between-data-analytics-and-ai-machine-learning/#47972eaa5878. Accessed 2019 Dec 09.
6. An overview of clinical applications of artificial intelligence. *(CADTH issues in emerging health technologies no. 174)*. Ottawa (ON): CADTH; 2018: https://www.cadth.ca/sites/default/files/pdf/eh0070_overview_clinical_applications_of_AI.pdf. Accessed 2019 Dec 11.
7. Gao S, Calhoun VD, Sui J. Machine learning in major depression: from classification to treatment outcome prediction. *CNS Neurosci Ther.* 2018;24(11):1037-1052.
8. An introduction to artificial intelligence and machine learning in addiction & mental health. Edmonton (AB): Alberta Health Services; 2018: https://www.albertahealthservices.ca/assets/info/res/mhr/if-res-mhr-ai-machine-learning.pdf. Accessed 2019 Dec 06.
9. Wells C, Hill S, Argáez C. Artificial intelligence and machine learning in mental health services: an environmental scan. *A joint publication with MHCC.* Ottawa (ON): CADTH; 2019 Dec. Accessed 2019 Dec 06.
10. Achtyes ED, Halstead S, Smart L, et al. Validation of Computerized Adaptive Testing in an Outpatient Nonacademic Setting: The VOCATIONS Trial. *Psychiatr Serv.* 2015;66(10):1091-1096.
11. Eichstaedt JC. Predicting and characterizing the health of individuals and communities through language analysis of social media. *Dissertation Abstracts International: Section B: The Sciences and Engineering.* 2018;79(1-B(E)).
12. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. *Curr Opin Psychol.* 2016;9:77-82.
13. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc.* 2018;25(9):1248-1258.
14. Narita S, Ohtani N, Waga C, Ohta M, Ishigooka J, Iwahashi K. A pet-type robot artificial intelligence robot-assisted therapy for a patient with schizophrenia. *Asia Pac Psychiatry.* 2016;8(4):312-313.
15. Scoglio AA, Reilly ED, Gorman JA, Drebing CE. Use of Social Robots in Mental Health and Well-Being Research: Systematic Review. *J Med Internet Res.* 2019;21(7):e13322.
16. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ.* 2017;358:j4008. http://www.bmj.com/content/bmj/358/bmj.j4008.full.pdf. Accessed 2019 Dec 09.
17. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health.* 1998;52(6):377-384. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1756728/pdf/v052p00377.pdf. Accessed 2019 Dec 09.
18. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529-536.
19. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol.* 2009;62(10):e1-e34.
20. Colombo D, Fernandez-Alvarez J, Patane A, et al. Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: a systematic review. *J Clin Med.* 2019;8(4):05.
21. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. *J Affect Disord.* 2019;245:869-884.
22. Lee Y, Ragguett RM, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J Affect Disord.* 2018;241:519-532.
23. Passos IC, Ballester P, Barros RC, et al. 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 Disord.* 2019;29:29.
24. de Filippis R, Carbone EA, Gaetano R, et al. Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review. *Neuropsychiatr Dis Treat.* 2019;15:1605-1627.
25. Wongkoblap A, Vadillo MA, Curcin V. Researching mental health disorders in the era of social media: systematic review. *J Med Internet Res.* 2017;19(6):e228.
26. 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.

27. Breen MS, Thomas KGF, Baldwin DS, Lipinska G. Modelling PTSD diagnosis using sleep, memory, and adrenergic metabolites: an exploratory machine-learning study. *Hum Psychopharmacol,.* 2019;34(2):e2691.

28. Byun S, Kim AY, Jang EH, et al. Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol. *Comput Biol Med.* 2019;112:103381.

29. Byun S, Kim AY, Jang EH, et al. Entropy analysis of heart rate variability and its application to recognize major depressive disorder: a pilot study. *Technol Health Care.* 2019;27(S1):407-424.

30. Deng Y, Hung KSY, Lui SSY, et al. Tractography-based classification in distinguishing patients with first-episode schizophrenia from healthy individuals. *Prog Neuropsychopharmacol Biol Psychiatry.* 2019;88:66-73.

31. Devine J, Fliege H, Kocalevent R, Mierke A, Klapp BF, Rose M. Evaluation of computerized adaptive tests (CATs) for longitudinal monitoring of depression, anxiety, and stress reactions. *J Affect Disord.* 2016;190:846-853.

32. Ding X, Yue X, Zheng R, Bi C, Li D, Yao G. Classifying major depression patients and healthy controls using EEG, eye tracking and galvanic skin response data. *J Affect Disord.* 2019;251:156-161.

33. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Ment Health.* 2018;5(4):e64.

34. He Q, Veldkamp BP, Glas CA, de Vries T. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment.* 2017;24(2):157-172.

35. Jaroszewski AC, Morris RR, Nock MK. Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *J Consult Clin Psychol.* 2019;87(4):370-379.

36. Jimenez-Serrano S, Tortajada S, Garcia-Gomez JM. A Mobile health application to predict postpartum depression based on machine learning. *Telemed J E Health.* 2015;21(7):567-574.

37. Kalmady SV, Greiner R, Agrawal R, et al. Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *NPJ Schizophr.* 2019;5(1):2.

38. Leightley D, Williamson V, Darby J, Fear NT. Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort. *J Ment Health.* 2019;28(1):34-41.

39. McGinnis RS, McGinnis EW, Hruschak J, et al. Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning. *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine & Biology Society.* 2018;2018:3983-3986.

40. McGinnis EW, Anderau SP, Hruschak J, et al. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE J Biomed Health Inform.* 2019;26:26.

41. Mohr DC, Schueller SM, Tomasino KN, et al. Comparison of the effects of coaching and receipt of App recommendations on depression, anxiety, and engagement in the IntelliCare Platform: factorial randomized controlled trial. *J Med Internet Res.* 2019;21(8):e13609.

42. Oh J, Yun K, Maoz U, Kim TS, Chae JH. Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm. *J Affect Disord.* 2019;257:623-631.

43. Oh K, Kim W, Shen G, et al. Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization. *Schizophr Res.* 2019;05:05.

44. Ramkiran S, Sharma A, Rao NP. Resting-state anticorrelated networks in Schizophrenia. *Psychiatry Res Neuroimaging.* 2019;284:1-8.

45. Schwarz E, Doan NT, Pergola G, et al. Reproducible grey matter patterns index a multivariate, global alteration of brain structure in schizophrenia and bipolar disorder. *Transl Psychiatry.* 2019;9(1):12.

46. Walsh-Messinger J, Jiang H, Lee H, Rothman K, Ahn H, Malaspina D. Relative importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning. *Psychiatry Res,.* 2019;278:27-34.

47. Wang Y, Iyengar V, Hu J, et al. Predicting future high-cost schizophrenia patients using high-dimensional administrative data. *Front Psychiatr.* 2017;8:114.

48. Wang S, Pathak J, Zhang Y. Using electronic health records and machine learning to predict postpartum depression. *Stud Health Technol Inform.* 2019;264:888-892.

49. Zhao N, Zhang Z, Wang Y, et al. See your mental state from your walk: recognizing anxiety and depression through Kinect-recorded gait data. *PLoS ONE [Electronic Resource].* 2019;14(5):e0216591.

50. Zhou L, Baughman AW, Lei VJ, et al. Identifying patients with depression using free-text clinical documents. *Stud Health Technol Inform.* 2015;216:629-633.

51. Zhuang H, Liu R, Wu C, et al. Multimodal classification of drug-naive first-episode schizophrenia combining anatomical, diffusion and resting state functional resonance imaging. *Neurosci Lett.* 2019;705:87-93.

52. Hilbert K, Lueken U, Muehlhan M, Beesdo-Baum K. Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: a multimodal machine learning study. *Brain Behav.* 2017;7(3):e00633.

53. Statistics Canada. Section G - Schizophrenia. Ottawa (ON): Government of Canada; 2015 Nov: https://www150.statcan.gc.ca/n1/pub/82-619-m/2012004/sections/sectiong-eng.htm. Accessed 2019 Dec 09.

54. Burns MN, Begale M, Duffecy J, et al. Harnessing context sensing to develop a mobile intervention for depression. *J Med Internet Res.* 2011;13(3):e55-e55.

55. Standing Senate Committee on Social Affairs Science and Technology, Ogilvie K, Eggleton A. Challenge ahead: integrating robotics, artificial intelligence and 3D printing technologies into Canada's healthcare systems. Ottawa (ON): Senate of Canada; 2017: http://publications.gc.ca/collections/collection_2017/sen/yc17-0/YC17-0-421-18-eng.pdf. Accessed 2019 Dec 09.

56. Gibbons RD, Weiss DJ, Frank E, Kupfer D. Computerized adaptive diagnosis and testing of mental health disorders. *Annu Rev Clin Psychol.* 2016;12:83-104.

57. Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci Biobehav Rev.* 2017;74(Pt A):58-75.

58. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. *Nat Lang Eng.* 2017;23(5):649-685.

59. Science Direct. Random forest. 2019; https://www.sciencedirect.com/topics/engineering/random-forest Accessed 2019 Dec 09.

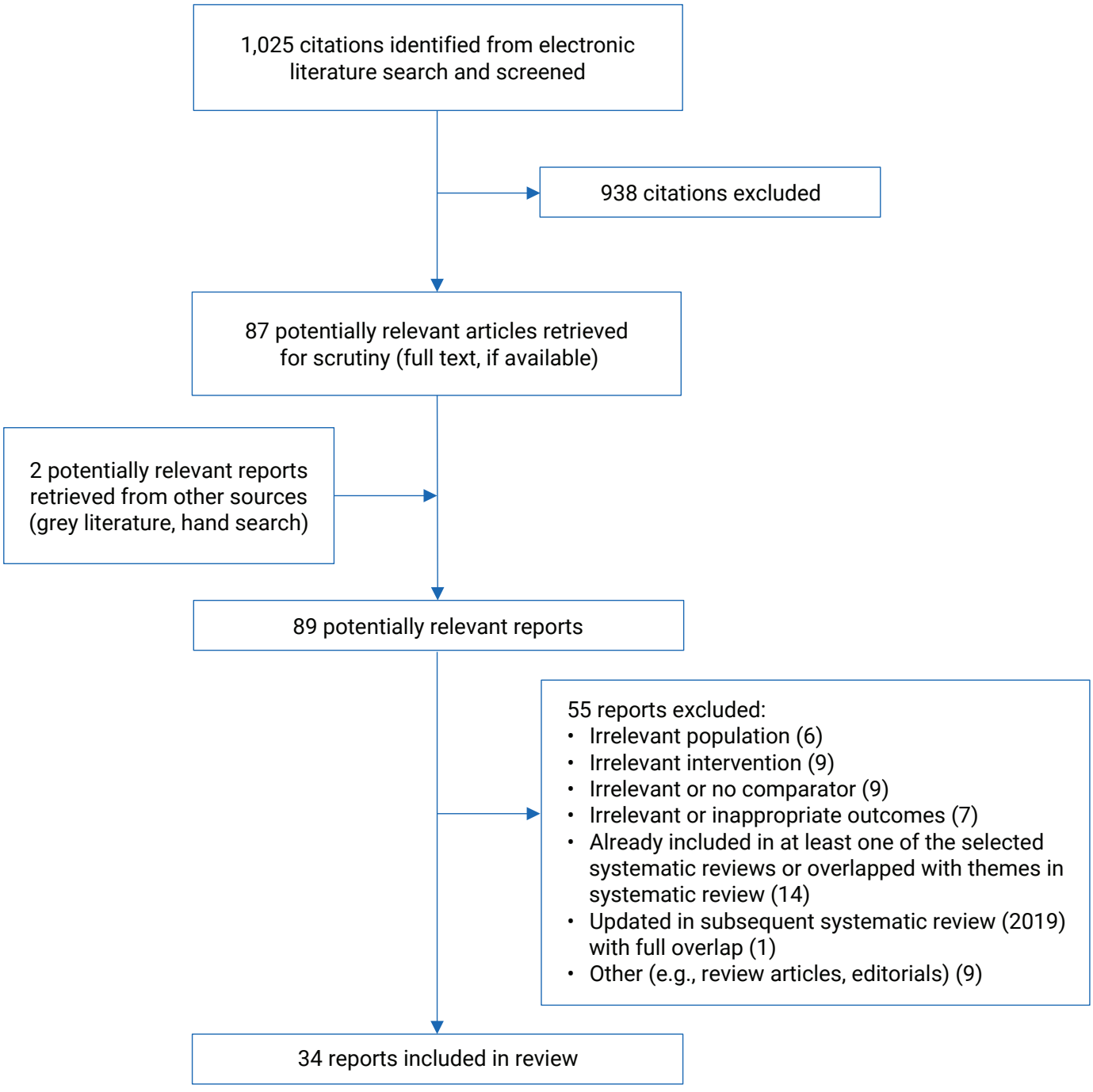
60. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol.* 2018;14:91-118.

Appendix 1: Glossary of Artificial Intelligence and Diagnostic Accuracy Terms

Table 2: Glossary of Artificial Intelligence and Diagnostic Accuracy Terms

Term	Definition
Area under the curve (AUC)	AUC is a probability value that ranges from 0 to 1. It is an aggregate measure of performance of a test over varying "thresholds" for success. An AUC of 1 represents an algorithm or test with 100% of its classifications being correctly classified. An AUC of 0.5 represents an algorithm or test that performed no better than chance.
Artificial intelligence	The reproduction of human cognition (i.e., reasoning, thinking, understanding) through an artificial means such as a computer.[55]
Artificial neural network	A form of artificial intelligence designed with neural "layers" — one input layer of neurons, multiple "hidden" layers of neurons, and a final output layer.[8]
Chatbot	Artificial intelligence–driven conversational agent programs that have the ability to "talk" with participants.[13] Examples of chatbots include customer service chatbots (e.g., "LiveChat," Facebook Messenger–based chatbots)
Computerized adaptive testing	An adaptive testing method that is targeted to the particular individual taking the test. The test gauges the estimate of whatever the targeted output is based on the previous answers to questions and adapts the test to provide more informative questions.[56]
Conversational agent	A system that mimics human conversation through text or spoken language.[13] These conversational agents include chatbots and "assistants" such as Siri, Alexa, and Google Home.
Convolutional neural network	An artificial neural network originally designed for images, with an input and output layer as well as a convolutional layer, pooling layer, and fully connected layer.[57]
Cross validation	A statistical method to generalize training of the algorithm.[7] The data set is split into a training set and a testing set, either simply (via hold-out sets) or through other techniques such as k-fold and leave-one-out.[7] It is a method to test how well the model will generalize to other datasets.
Deep learning	The layers within a neural network.[8]
Machine learning	Algorithms that "learn" from data to generate outputs rather than those that are programmed to deliver a fixed solution.[7]
Natural language processing	A machine learning technique in which inferences are made from text or speech about the speaker's thoughts, feelings, and motivations.[58]
Random forest	An ensemble learning method of multiple independent decision trees. Each tree casts a vote for a particular output, the output with the majority of votes is the final determined output.[59] This is similar to casting a vote in an election in which whatever had the majority of votes is the "winning" output.
Sensitivity	Also known as the true-positive rate or recall value, sensitivity is the proportion of individuals with a disease or condition who are correctly identified by the algorithm or test as having the disease or condition.
Specificity	Also known as the true-negative rate, specificity is the proportion of individuals without a disease or condition who are correctly identified by the algorithm or test as not having the disease or condition.
Supervised learning	Supervised machine learning involves labelling the cases during the training of the data set so that the outputs (classified groups) already have assigned names.[60] For example, if you are labeling a picture as "cat" or a "dog" when the model was initially trained, all of the data used in training was labeled by humans as "cat" or "dog" so the algorithm could learn from it.
Support vector machines	A linear classification method in which a hyperplane is drawn between two classes based on the maximal distance from two support vectors (data points).[3] Kernels can be used to transform data to further discriminate the two classes.[48]
Unsupervised machine learning	Unsupervised learning is when cases are unlabeled and the machine learning algorithm divides the sample in groups of related cases, with no assigned names for the outputs.[60] The data are categorized based on the properties of the data itself.[8] In the "cat" or "dog" picture example, when the model was trained on previous data, this data was not already labeled as "cat" or "dog."

Appendix 2: Selection of Included Studies



Appendix 3: Characteristics of Included Publications

Table 3: Characteristics of Included Systematic Reviews and Meta-Analyses

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
Burke (2019)[21] US University of Michigan James N. Morgan Fund grant	All study types (peer-reviewed) 35 included studies Studies examining: Suicide death, n = 5 Suicide attempt, n = 14 Suicide planning, n = 1 Suicidal ideation, n = 10 Suicide risk, n = 8 Non-suicidal self-injury, n = 3 Through February 2018	Included patients who had one or more of the outcomes of non-suicidal self-injury, suicidal ideation, suicide planning, suicide attempt, and suicide death This included patients with a variety of mental health conditions; e.g., bipolar, MDD, mood disorders, history of self-injury, SCZ, personality disorders, suicidal ideation This also included adolescents, adults, undergraduate students, and older adults	ML techniques for the prediction of suicide-related events Included all types of machine learning, but only identified studies using SML — including regularized regression, decision trees (interpretable nonlinear methods), random forests, or boosting (less interpretable, or "black box" methods)	NR	Interpretable nonlinear methods vs. other less interpretable methods ML compared with "traditional methods;" e.g., logistic regression	Accuracy (sensitivity, specificity, AUC, R2, NPV, PPV, recall value)

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
Colombo (2019)[20] Spain Marie Curie EF-ST AffecTech	All study types, except reviews and systematic reviews, meta-analyses, case reports, editorials, and other editorial materials 40 included studies 1 relevant study March 2019 (Note: unknown what date limitations were in place)	Included patients with primary (both past and current) MDD diagnosed with the criteria from the DSM	Smartphone-based or handhold technology – based EMA and EMI[a] that collected daily self-reports, not including paper-and-pencil–based EMA. EMI was provided through handheld technologies as a stand-alone or combined intervention. This included EMI data collected through wearable sensors or device sensors. Included ML-related EMI was "Mobylize!," "a context-aware system, composed of three main elements: (1) A mobile application for the collection of self-reports; (2) a website with feedback and theoretical lessons; (3) periodic contacts with trained coaches" (P. 9), used for 8 weeks 5 times a day. The machine learning algorithm predicts the state of the patient (mood, context, activities, etc.), and sends tailored feedback to participants	NR	Mobylize! vs. no comparator (single-arm field trial)	Depressive symptoms, accuracy

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
de Filippis (2019)[24] Italy Funding NR	Studies with Jadad score > 3, with control groups 35 studies included 8 studies used sMRI, 26 used fMRI, 1 study used both fMRI and sMRI to determine patients with SCZ from those with other psychiatric disorders or healthy controls Up to December 2018	Patients with SCZ (diagnosed with *DSM-IV*, *DSM-IV-TR*, *DSM-V*, or ICD-10 chronic or newly diagnosed), any episode number, either on or not on antipsychotic drugs	ML techniques for differentiating patients with SCZ and healthy controls (i.e., diagnostics) using neuroimaging (either fMRI or sMRI) sMRI ML techniques included ridge, LASSO, elastic net and L0 Norm regularized logistic regression, support vector classifier, regularized discriminant analysis, RF and a Gaussian process classifier, RFE fMRI ML techniques included MVPA, FC density analysis, SVM, LASSO, GSM, DGM, ROCA, ELM, DNN, leave-one-out SVM, voxel-mirrored homotopic connectivity, sparse autoencoder network, DANS, translation-based multimodal fusion approach	Individuals involved in diagnostics (e.g., clinicians)	ML methods vs. diagnostic methods not specified	Accuracy (sensitivity, specificity, AUC, precision, error rate)

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
Passos (2019)[23] Canada Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); CAPES (Brazilian Government); FIPE (Hospital de Clínicas de Porto Alegre); Canadian Institutes of Health Research, Grant/ Award Number: 103703, 106469 and 142255; Nova Scotia Health Research Foundation; Dalhousie Clinical Research Scholarship; Brain & Behavior Research Foundation; 2007 Young Investigator and 2015 Independent Investigator Awards; Ministry of Health, Grant/Award Number: 16-32791A and 16-32696A; Stanley Medical Research Institute; CAPES (Brazilian Government)	90 studies were included at screening and 1 study was included through reference review Eligible designs NR N = 17 classification studies using structural neuroimaging and DTI N = 19 classification studies using functional neuroimaging N = 6 classification studies using genetic analysis N = 6 classification studies using electroencephalografic measures N = 6 classification studies using neuropsychological tests and mood symptoms N = 5 classification studies using blood biomarkers N = 3 classification studies using standard sensors N = 1 classification studies using text N = 19 machine learning studies predicting clinical outcomes of bipolar disease (depression relapse, severity, mood changes, suicide, quality of life, aging, hyper-reactivity, other) N = 5 machine learning studies predicting treatment response and adverse effects N = 5 machine learning studies using unsupervised or semi-supervised algorithms on BPD Articles met the inclusion criteria if they assessed patients with BPD using machine learning techniques Published between January 1960 and January 2019	Adult patients (older than 18 years) with BPD Compared with SCZ, unipolar depression, healthy controls, and other conditions (not specified)	Various ML “techniques” How machine learning and big data will contribute to studying BPD in improving outcome predictions in prevention, diagnosis and treatment	NR	ML methods vs.diagnostic methods not specified	Diagnostic accuracy (accuracy, sensitivity, specificity, AUC, true-positive, false-positive, true-negative and false-negative)

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
Gao (2018)[7] China National High-Tech Development Plan (863), Grant/Award Number: 2015AA020513; NIH Grant, Grant/ Award Number: 1R01MH094524, P20GM103472 and R01EB005846; Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/ Award Number: XDBS01000000; "100 Talents Plan" of Chinese Academy of Sciences, the Chinese Natural Science Foundation, Grant/ Award Number: 61773380 and 81471367	66 included studies Eligible designs NR Published between January 2000 and December 2017	Patients with MDD	Machine learning methods for the classification of patients with MDD using MRI	Not specified	ML methods vs. diagnostic methods not specified	Accuracy, sensitivity, specificity

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
Laranjo (2018)[13] Australia National Health and Medical Research Council grant and Program Grant	17 included studies Studies with evaluations of human interactions with the system, excluding "wizard of Oz" studies 5 relevant studies Through April 2017, and updated in February 2018	Participants with mental health conditions (e.g., depression, anxiety, PTSD) Participants with autism spectrum disorder	Conversational agents that use any unconstrained natural language input, with which humans can interact on a turn-by-turn basis; allowing for more than one turn for the human	Consumers, caregivers, or health care professionals	Educational eBooks, psychiatrist visit, or written and audio content	Depression symptoms, anxiety symptoms, meditation frequency, acceptability, PTSD symptoms Diagnostic accuracy

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
Lee (2018)[22] Canada Funding NR	26 studies included in qualitative synthesis; 20 studies included in quantitative synthesis 16 studies investigated predictors of response with a pharmacological intervention; 1 study investigated psychotherapy; 2 studies combined antidepressants and psychotherapy; 7 studies investigated neuromodulatory treatment (e.g., rTMS, ECT, tDCS). Inception to February 8, 2018	Adults (older than 18 years) diagnosed with bipolar or unipolar depression, defined with a diagnostic manual (i.e., DSM ICD) Treatment responders vs. non-responders	Predictors (neuroimaging, phenomenological, genetic or combined) for prediction of treatment response using SML and classification algorithms (n = 24) or unsupervised learning method (n = 2) Commonly used models were linear kernel-based SVM, L1-regularized logistic regression, logistic regression with elastic net regularization, and linear artificial neural networks, radial basis function kernel-based SVM, alternating or hierarchical multi-label decision trees, and multi-layer perceptron ANN Binary (n = 22) or non-binary (n = 2) classification system in SML, multivariate classifiers (n = 1) and hierarchical clustering algorithms (n =1) Treatments were evidence-based and guideline-concordant treatments for depression (e.g., neuromodulation, pharmacotherapy, psychotherapy)	Not reported	ML methods vs. diagnostic methods not specified	Depression-related outcomes (e.g., mood symptom severity, occupational or psychosocial functioning, depression-related hospital admission frequency or duration, suicidal ideation) Classification accuracy (percentage rate, ROC, AUC)

First Author (Publication Year) Country, Funding	Study Designs and Numbers of Primary Studies Included, Search Date Range	Population Characteristics	Intervention and Purpose	Intended Users	Comparator(s)	Clinical Outcomes
Wongkoblap (2017)[25] UK UK National Institute for Health Research Biomedical Research Centre – Based at Guy's and St Thomas' NHS Foundation Trust and King's College London	48 included studies Peer-reviewed studies 2010 to March 2017	Mental health problems as defined by the UK National Institute for Health and Care Excellence Included studies examined depression, postpartum depression, PTSD, anxiety, OCD, borderline personality disorder, BPD, SAD, eating disorders, SCZ, ADHD, sleep disorder, and suicidality	Prediction or classification models using ML techniques that used social media (text posts, network interactions, or other features)	NR	Social media content vs. questionnaires (e.g., PHQ-9, Beck Depression Inventory, Zung Self-Rating Depression Scale, Depressive Symptom Inventory-Suicidality Subscale, Symptom Checklist-90 Revised, Suicide Probability Scale, Acquired Capability for Suicide Scale, Interpersonal Needs Questionnaire, PNAS, Psychological Well-Being Scale)	Accuracy Prediction of mental illness

ADHD = attention-deficit/hyperactivity disorder; ANN = artificial neural network; AUC = area under the curve; BPD = bipolar disorder; DANS = discriminant autoencoder network with sparsity constraint; DGM = deep neural generative model; DNN = deep neural network; DSM = *Diagnostics and Statistics Manual of Mental Disorders; DSM-IV = Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition*; *DSM-IV-TR = Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition, Text Revision; DSM-V = Diagnostics and Statistics Manual of Mental Disorders, Fifth Edition*; DTI = diffusion tension imaging; ECT = electroconvulsive therapy;
ELM = extreme learning machine; EMA = ecological momentary assessment; EMI = ecological momentary interventions; FC = functional connectivity; fMRI = functional magnetic resonance imaging; GSM = generalized sparse model; HC = healthy controls; ICD = International Statistical Classification of Diseases and Related Health Problems; ICD-10 = International Statistical Classification of Diseases and Related Health Problems, 10th Revision; LASSO = least absolute shrinkage and selection operator; MDD = major depressive disorder; ML = machine learning; MRI = magnetic resonance imaging; MVPA = multivariate pattern analysis; NHS = National Health Service;
NIH = National Institute of Health; NPV = negative predictive value; NR = not reported; OCD = obsessive compulsive disorder; PHQ-9 = Patient Health Questionnaire-9; PNAS = Proceedings of the National Academy of Sciences; PPV = positive predictive value; PTSD = post-traumatic stress disorder; RF = random forest; RFE = recursive feature elimination; ROC = receiver operating characteristic; ROCA = receiver operating characteristic curve analysis; rTMS = repetitive transcranial magnetic stimulation; SAD = seasonal affective disorder; SCZ = schizophrenia; SML = supervised machine learning; sMRI = structural magnetic resonance imaging; SVM = support vector machine; tDCS = transcranial direct current stimulation; vs. = versus.

[a] EMA is the sampling of behaviours, thoughts, and experiences in real time, within the "natural" environment or context. EMI is providing extended treatment that occurs in a real-life context, outside of clinical settings. Not all EMAs and EMIs are AI or machine learning based.

Table 4: Characteristics of Included Primary Clinical Studies

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Jaroszewski 2019[35] US Chet and Will Griswold Suicide Prevention Fund and the For the Love of Travis Fund[a]	Randomized controlled trial	Users of the Koko application between August 10, 2017 and September 20, 2017 N = 39,450	Develop and evaluate a brief, automated risk assessment and intervention platform (digital mental health app Koko – a psychoeducation intervention designed to reduce perceived barriers to the use of crisis resources) Koko is a text-based user interface (first-person persona – "KokoBot") available on messaging services, desktop and mobile browsers. It is supervised by a machine learning algorithm that evaluates the semantic content of posts using recurrent neural networks and word embeddings. If confidence in post is classified as a crisis post with 0.95 confidence, it is automatically ruled as crisis, if below 0.95, it is reviewed by 1 of 3 human moderators. If classified as crisis post, Koko messages user based on the National Suicide Prevention Lifeline, if user assessed as high risk, they were presented with crisis resources Meant to increase use of crisis resources for individual experiencing a mental health crisis	Varying versions of the application Intervention version - shown a list of crisis resources relevant to their issue and country of origin (e.g., the National Suicide Prevention Lifeline) then received additional interventionsb Randomized n = 19,612 treatment n = 19,838 control Allocated and received intervention or control (post classified as crisis and posted on network) Treatment, n = 775 Control, n = 805 Analyzed Treatment, n = 325 Control, n = 327	Patients (peer to peer with no clinical or counsellor oversight)	Accuracy of assessment Use of crisis resources Satisfaction with service 5-hour follow-up

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Breen, 2019[27] US National Research Foundation and the A.W. Mellon Foundation	Cross-sectional quasi-experimental design	The study population consisted of three groups: 1. Women with a diagnosis of PTSD (the diagnosis was made using a 45-point cut-off on the CAPS) 2. Women who experienced trauma but did not have PTSD 3. Women without a history of trauma or PTSD (healthy controls) N = 60 (20 per group)	Using SVM to determine PTSD using sleep, cognitive, and biochemical variables	SVM vs. Clinician-Administered PTSD Scale	Clinicians	Sensitivity Specificity Accuracy Follow-up NA
Byun, 2019a[28] South Korea Funded by a series of grants from the Ministry of Science and ICT of the South Korea government.	Diagnostic cross-sectional study with case-control selection	Participants with MDD were matched with a cohort of healthy controls (by age and gender). MDD was diagnosed by a board-certified psychiatrist based on *DSM-IV* criteria. N = 78 (37 diagnosed with MDD; 41 healthy controls)	Using HRV analysis and SVM algorithms to determine presence of MDD	SVM vs. board-certified psychiatrist diagnosis using *DSM-IV*	Clinicians	Sensitivity Specificity PPV NPV Accuracy Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Byun, 2019b[29] South Korea National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1C1B5017730) and the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2015-0-00062, Original Technology Research Program for Brain Science through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. NRF-2016M3C7A1947307; PI HJJ), and the Bio and Medical Technology Development Program of the NRF funded by the Korean government, MSIP (No. NRF-2017M3A9F1027323; PI HJJ).	Diagnostic cross-sectional study with case-control selection	MDD patients and age and gender matched healthy controls that were recruited from the Samsung Medical Center, Seoul, Korea. N = 33 MDD; N = 33 HC	SVM-RFE and four ML algorithms (neuro-fuzzy networks, linear discrimination analysis (LDA), LR, and Bayesian networks) to determine MDD from entropy analysis of HRV	SVM-RFE vs. senior psychiatrist evaluation	Clinicians (diagnostics)	Accuracy Sensitivity Specificity PPV NPV Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Deng, 2019[30] China Support was received from the Philip KH Wong Foundation, the Beijing Training Project for Leading Talents in S&T, a grant from the "Strategic Priority Research Program (B)" of the Chinese Academy of Sciences, and the CAS/SAFEA International Partnership Programme for Creative Research Teams.	Diagnostic cross-sectional study with case-control selection	Participants (between 18 and 40 years of age) with first-episode *DSM-IV* SCZ and schizoaffective disorder were matched with a cohort of healthy controls (by age, gender, and handedness). N = 125 (65 with first-episode SCZ; 60 healthy controls)	Determine which MRI identified features (tractography) were most important for the differentiation of individuals with first-episode SCZ from healthy controls using the RF model	RF vs. structured clinical interview performed by qualified psychiatrists	Clinicians	Diagnostic accuracy (e.g., sensitivity, specificity, NPV, PPV, overall accuracy) Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Ding 2019[32] China Financial support was provided by Peking University Sixth Hospital	Diagnostic cross-sectional study with case-control selection	Outpatients (aged between 18 and 60) with MDD were matched with a cohort of health controls on age, gender, and educational background. MDD diagnoses were determined by experienced psychiatrists. Exclusion criteria included a history of SCZ, mania, alcohol and drug abuse, or other mental health conditions, severe cardiovascular disease or other somatic disease that may affect visual or auditory functions, having received electroconvulsive therapy within one month, active suicidal intention, and having taken medicines which may significantly affect brain functions (e.g., clozapine or chlorpromazine). N = 348 (144 with MDD; 204 healthy controls)	Physiological data (i.e., electroencephalography, eye-tracking information, and galvanic skin response) to differentiate MDD from healthy controls, with Ml models such as RF, LR, and SVM	ML models vs. ICD-10 criteria	Clinicians	Accuracy, precision, recall sensitivity, and f1 scores for each machine learning algorithm Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Kalmady 2019[37] Canada IBM Alberta Centre for Advanced Studies and MITACS (IT09558) funds to S.V.K.; Wellcome Trust/DBT India Alliance (500236/Z/11/Z) and DST (DST/SJF/LSA-02/2014-15) research grants to G.V.; Alberta Machine Intelligence Institute and NSERC grants to R.G. V.S. is supported by the ICMR.	Diagnostic cross-sectional study with case-control selection	Antipsychotic-naive SCZ patients with age and gender matched healthy controls N = 81 (SCZ); N = 93 (HC)	An ensemble ML model for predicting SCZ called EMPaSchiz using fMRI data	ML vs. *DSM-IV* criteria for SCZ using Mini International Neuropsychiatric Interview (MINI) Plus	Clinicians (prediction)	Prediction accuracy (Accuracy, precision, sensitivity, specificity, true-positive, true-negative, false-positive, false-negative) Follow-up NA
Leightley 2019[38] UK Funding NR	Diagnostic accuracy cross-sectional study with cohort selection	Data were collected using the KCMHR longitudinal cohort N = 13,690	Supervised ML classifiers to predict probable PTSD using a UK military cohort ML classifiers included SVM, RF, ANN and Bagging	ML vs. PTSD civilian checklist	Clinician (diagnostics)	Sensitivity Specificity Matthews Correlation Coefficient Follow-up NA
McGinnis 2019[40] US NIMH Grant K23-MH080147, Michigan Institute for Clinical and Health Research , Blue Cross Blue Shield of Michigan Foundation Grant, Brain Behaviour Research Foundation, NIMH Grant R03MH102648	Prospective diagnostic test accuracy cohort study	Children aged 3 to 7 years without a suspected or diagnosed developmental disorder, serious medical condition, or taking medication that affect the central nervous system N = 71 63% female Mean age = 5.25	ML speech analysis (LR, SVM with linear kernel, LR with gaussian kernel, RF) of child voice recordings during a 3-minute speech taskc to detect anxiety and depressive symptoms or internalizing disorder	ML algorithm vs. structured clinical interview	Diagnostics (i.e., clinicians)	Accuracy Sensitivity Specificity ROC AUC Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Mohr 2019[41] US United States National Institute of Mental Health grant R01 MH100482 and research grant K08 MH112878 from the National Institute of Mental Health	Factorial RCT	Adults (18 year or older) patients with depression (PHQ-9 ≥ 10) or anxiety (GAD-7 ≥ 8) N = 301	IntelliCare platform – 13 apps (12 designed for a specific behavioural or psychological treatment strategy) via a mobile phone. Treatment is created using ML	2x2 factorial design: coaching vs. weekly reminders vs. weekly app recommendations vs. no recommendations	Patients	Depressive symptoms (PHQ-9, range 0 to 27) Anxiety (GAD-7, scale of 0 to 21) Engagement with app (time to last use, number of app sessions, and number of apps downloaded) Follow-up week 4, week 8 (end of treatment), 3 months, and 6 months
Oh 2019a[42] South Korea Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HM15C1054). Uri Maoz was funded by the Bial Foundation (grant 388/14).	Cross-sectional diagnostic accuracy study with case-control selection	Data were collected from the NHANES and K-NHANES to train DL algorithms and other ML classifiers. Data set from NHANES = 28 280 participants Data set from K-NHANES = 4,949 participants	Assess the utility of ML and DL in deciphering risk factors for depression. DL algorithm	ML vs. PHQ-9	Clinicians	AUC Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Oh 2019b[43] South Korea Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) and the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C2383), research funds of Chonbuk National University in 2018 and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant number: 2018R1A6A3A01013251).	Cross-sectional diagnostic accuracy study with case-control selection	Participants included individuals that met *DSM-IV-TR* criteria for SSDs (SCZ, schizoaffective disorder, and schizophreniform disorder) and matched health controls. N = 103 (SSD); N = 41 (HC)	CAE based CNN and SVM methods to distinguish SSD from HC CAE was also compared to other models	ML vs. structured clinical interview for *DSM-IV*	Clinicians (diagnostics)	Accuracy, AUC, sensitivity, specificity, PPV and NPV Follow-up NA
Ramkiran 2019[44] India Center of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 from the National Institutes of Health	Cross-sectional diagnostic study	Patients from the Mind Research Network, Centre for Biomedical Research and Excellence with SCZ and healthy controls aged 18 to 65 years. N = 56 (SCZ); N = 56 (HC)	SVM to discriminate HC from SCZ patients using anticorrelated networks	SVM vs. structured clinical interview for *DSM-IV*	Clinicians (diagnostics)	Accuracy Sensitivity Specificity Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Schwarz 2019[45] Germany European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602450 (IMAGEMEND, IMAging GENetics for MENtal Disorders) and the Deutsche Forschungsgemeinschaft (DFG), SCHW 1768/1-1. A.M.-L. was supported by the Deutsche Forschungsgemeinschaft (DFG) (Collaborative Research Center SFB 636, subproject B7); the German Federal Ministry of Education and Research (BMBF) through the Integrated Network IntegraMent (Integrated Understanding of Causes and Mechanisms in Mental Disorders) under the auspices of the e:Med Programme (BMBF Grant 01ZX1314A and 01ZX1314G); and the Innovative Medicines Initiative Joint Undertaking (IMI) under Grant Agreements no 115300 (European Autism Interventions—A Multicentre Study for Developing New Medications) and no 602805 (European Union- Aggressotype).	Diagnostic cross-sectional study with cohort selection method	Eight cohorts consisting of patients with SCZ (cohort I-IV), BPD (cohort V and VI), ADHD (cohort V to VI) and healthy controls (cohort I-VIII). N = 2,668	RF and SVM to differentiate patients with SCZ from controls and other disorders (BPD and ADHD), and to identify brain structures for successful classification. .	ML vs. established diagnoses (*DSM-IV*)	Clinicians (diagnostics)	Classification accuracy (AUC, sensitivity, specificity) Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Walsh-Messinger 201946 US Hongshik Ahn: Ministry of Science, ICT and Future Planning, Korea, under the "ICT Consilience Creative Program" (IITP-2017-R0346-15- 1007) supervised by the Institute for Information & Communications Technology Promotion. Dolores Malaspina: NIMH R01MH06642; NIMH RC1MH088843; NIMH K24MH00169.	Cross-sectional diagnostic accuracy study with case-control selection	Data were collected from psychiatric cases and health comparison participants between 1995 and 2010. N = 113 psychiatric cases; N = 51 HC	RF to classify SCZ and other psychiatric diagnosis	ML vs. Diagnostic Interview for Genetic Studies and clinician assessment	Clinicians (disorder classification)	Accuracy No follow-up reported
Wang 2019[48] US Walsh McDermott Scholarship, R01 MH105384, P50 MH113838 and China Scholarship Council.	Diagnostic longitudinal prediction study	Pregnant women with singleton births at risk of PPD N = 9,980	Six ML models (L2-regularized LR, SVM, Decision Tree, Naive Bayes, XGBoost, and RF) Use of electronic health record data to predict PPD post-birth	ML vs. ICD-10 codes indicating PPD post-birth	Clinicians and individuals involved in postnatal care	Sensitivity Specificity AUC

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Zhao 2019[49] China National Key Research & Development Program of China (2016YFC1307200); and National Natural Science Foundation of China (31700984)	Cross-sectional diagnostic accuracy study with cohort selection	Graduate students from the University of Chinese Academy of Science N = 179	Kinect-recorded ML model for determining anxiety and depression levels through walking gait. ML models were trained using SLR, LR, e-SVR, n-SVR and GP to predict anxiety and depression scores. Participants completed the GAD-7 and PHQ-9 then walked on the footpath for two minutes with Kinect Cameras recording their gait	ML vs. GAD-7 and PHQ-9	Clinicians (diagnostic accuracy)	Precision, Recall F-measure
Zhuang 2019[51] China National Natural Science Foundation of China (Grant Nos. 81871083, 61671292, 6181101049 and 61375112), Foundation of Shanghai Jiao Tong University (YG2017ZD13)	Cross-sectional diagnostic accuracy study with case-control selection	Drug-naive FES patients and matched healthy controls N = 40 FES patients N = 29 HC	Diagnose FES patients from healthy controls using multi-kernel SVM and combined structural MRI, DTI, and resting state-fMRI	ML vs. clinician assessment (*DSM-IV*)	Clinicians (FES classification)	Classification accuracy

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Fulmer 2018[33] US X2 AI Inc. (US.)	RCT	Students attending University in the US. (nonclinical college population) Aged 18 years and older	Tess – AI chatbot Tess is a psychological AI chatbot that provides mental health support (CBT, mindfulness-based therapy, emotionally focused therapy, acceptance and commitment therapy, motivational interviewing, self-compassion therapy, and interpersonal psychotherapy) and psychoeducation through conversation.	4 weeks of Tess vs. 2 weeks of Tess vs. eBook on depression After the study period, participants were contacted to complete a second set of questionnaires online.	Patients Tess was designed to deliver personalized conversations based on the expressed emotions and mental health concerns of participants, not to replace trained therapists.	Symptoms of anxiety, depression (PHQ-9, GAD-7, PANAS), Engagement 2 weeks (Tess for 2 weeks group) or 4 weeks (education group or Tess for 4 weeks group)
McGinnis 2018[39] US NIMH (K23-MH080147, R03-MH102648), the Michigan Institute for Clinical and Health Research (UL1TR000433), the Blue Cross Blue Shield of Michigan Foundation Grant (1982. SAP), the Biomedical and Social Sciences Scholar Program, and the Brain Behavior Research Foundation.	Prospective diagnostic accuracy cohort study	N = 63 children (aged 3 to 7) and their primary caregivers	Machine learning (SVM with linear kernal, DT, and kNN) to diagnose internalizing disorders in children Single wearable sensor with a 90-second fear-induction task, with the best 20 seconds of data from the "Potential Threat phase" of the task extracted for analysis	ML vs. multimodal assessments (diagnostic interviews)	Clinicians	Accuracy Sensitivity Specificity AUC Follow-up NA

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
He 2017[34] US Stichting Achmea Slachtofferhulp Samenleving, the Netherlands.	Cross-sectional diagnostic accuracy study with case-control selection	Trauma survivors (half diagnosed with PTSD and half non-PTSD) N = 300	Use ML in PTSD screening and diagnosis from trauma survivors using narrative accounts and NLP	ML vs. diagnosis obtained by the practitioners via structured interviews with standardized instruments (DSM -IV and Clinician-Administered PTSD Scale)	Intended for screening for PTSD patients	Accuracy Sensitivity Specificity PPV NPV No follow-up mentioned
Wang 2017[47] US Funding: ODH, Inc.	Retrospective accuracy study	Patients with SCZ Training data set n = 34,510 Testing data set n = 30,077	Predictive model for identification of high-cost (health expense) SCZ patients Three model types: Baseline model (demographic data, total cost features) Enhanced model (coverage, health care utilization, antipsychotic medication usage + baseline model) Final model (sparse features + enhanced model)	Model vs. CMS-HCC model Model predicted cost vs. actual cost	Health organizations	R^2, PCA, and CA Follow-up of one year
Devine 2016[31] Germany Deutsche Forschungsgemeinschaft And Department of Psychosomatics and Psychotherapy, Charité University Hospital	Longitudinal retest-reliability study	Psychosomatic inpatients treated between 2007 and 2011 N = 595	3 CAT tests (IRT-techniques) D-CAT (depression) A-CAT (anxiety) S-CAT (stress)	CAT vs. conventional depression, stress, and anxiety instruments (PHQ-9, GAD-7, and PSQ)	Clinicians	Measurement precision, retest-reliability (between the initial interview and the admission to the hospital), sensitivity to change 96.5% of patients stayed 1 to 2 weeks, 60.2% stayed > 2 weeks 27.4% stayed > 3 weeks.

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Achtyes 2015[10] US Pine Rest Foundation – CAT-DI/SCID Assessment Tool and the National Institute of Mental Health – MH66302	Cross-sectional diagnostic accuracy study with case-control selection	Adults (18 to 70 years) presenting to mental health care clinic, excluding those with SCZ, schizoaffective disorder or other psychotic disorder, organic mood disorder due to a general medical condition or substance use disorder; drug or alcohol dependence in the prior 3 months, requiring in-patient hospitalization due to suicide risk or psychosis, or Alzheimer or Parkinson disease N = 145	CAT-MH test Diagnostic tests using CAT, with addition ML components such as decision trees and RF	CAT-MH (includes CAT-D, CAT-DI, CAT-ANX, CAT-MANIA) vs. conventional assessments (SCID, HAM-D_{25}, PHQ-9, CES-D, GAF)	Clinicians	Sensitivity, Specificity, Correlation with gold-standard symptom severity scales
Jimenez-Serrano 2015[36] Spain Spanish Ministerio de Sanidad (grant PIO41635, Vulnerabilidad gene´tico-ambiental a la depresio´n posparto, 2006–2008) and the Instituto de Salud Carlos III (RETICS Combiomed, grant RD07/0067/2001)	Diagnostic accuracy study with cohort selection	Postpartum women from seven Spanish general hospitals N = 1,397 women after 32-week follow-up after child birth	Use of four different types of PR classifiers: Naive Bayes, LR, SVM, and ANN PR models were intended to detect risk of PPD during the first week postpartum intervention. PR models could be inserted into an mHealth app with a CDSS for mothers and clinicians	PR models vs. other PR models PR model vs. Spanish EPDS test version and DIGS	Clinicians Postpartum mother to predict postpartum depression during the first week after childbirth	Accuracy Sensitivity Specificity AUC

First Author, Publication Year, Country, Funding Source	Study Design	Population Characteristics	Intervention and Purpose	Comparator(s)	Intended Users	Clinical Outcomes, Length of Follow-Up
Zhou 2015[50] US Partner Siemens Research Council	Retrospective diagnostic accuracy study	Patients with a history of ischemic heart disease N = 1200 Training set N = 600 Testing set = 600 Cases were classified as depression with high, intermediate and low confidence based on information in discharge summary	MTERMS NLP on free-text discharge summaries to determine depression status Use of NLP on unstructured narratives that will identify patients at high risk of hospital readmission ML techniques were SVM, generalized nearest neighbour classifier, Repeated Incremental Pruning to Produce Error Reduction rule learner, and a C4.5 DT learner	NLP approach vs. manual review by domain experts (pharmacy doctoral student in consultation with an internal medicine physician) Free-text NLP model vs. coded diagnoses	Clinicians	Recall (i.e., sensitivity) Precision (i.e., PPV) F-measure

ADHD = attention-deficit/hyperactivity disorder; AI = artificial intelligence; ANN = artificial neural network; AUC = area under the curve; BPD = bipolar disorder; CA = cost accuracy; CAE = 3-D convolutional autoencoder; CAPS = Clinical-Administered PTSD Scale; CAT = computerized adaptive test; CAT-ANX = computerized adaptive test for anxiety severity; CAT-D = computerized adaptive test for depression diagnosis; CAT-DI = computerized adaptive test for depression severity; CAT-MANIA = computerized adaptive test for manic/hypomanic symptom severity; CAT-MH = computerized adaptive test for mental health; CBT = cognitive behavioural therapy; CDSS = clinical decision support system; CES-D = Center for Epidemiologic Studies Depression Scale; CMS-HCC = Centers for Medicare & Medicaid Services Hierarchical Condition Categories; CNN = convolutional neural network; COBRE = Center of Biomedical Research Excellence; DIGS = diagnostic interview for genetic studies; DL = deep learning; *DSM-IV = Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition*; DT = decision tree; DTI = diffusion tension imaging; EPDS = Edinburgh Postnatal Depression Scale; e-SVR = Epsilon-SVR; FES = first-episode schizophrenia; fMRI = functional magnetic resonance imaging; GAD-7 = Generalized Anxiety Disorder 7-item scale; GAF = Global Assessment of Functioning; GP = gaussian processes; HAMD-D25 = Hamilton Rating Scale for Depression; HC = healthy controls; HRV = heart rate variability; ICD-10 = International Statistical Classification of Diseases and Related Health Problems 10th Revision; ICMR = Indian Council of Medical Research; ICT = information and communications technology; IITP = Institute for Information and Communications Technology Promotion; IRT = item response therapy; KCMHR = King's Centre for Military Health Research; KHIDI = Korean Health Industry Development Institute; K-NHANES = Korea–National Health and Nutrition Examination Survey; kNN = k-nearest neighbour; LDA = linear discriminate analysis; LR = logistic regression; MDD = major depressive disorder; MDD = major depressive disorder; MINI = Mini International Neuropsychiatric Interview; ML = machine learning; MRI = magnetic resonance imaging; MSIT = ministry of science and ICT; MTERMS = medical text extraction reasoning and mapping system; N = number; NA = not applicable; NHANES = National Health and Nutrition Examination Survey; NIMH = National Institute of Mental Health; NLP = natural language processing; NPV = negative predictive value; NR = Not Reported; NRF = National Research Foundation; NSERC = Natural Sciences and Engineering Research Council; n-SVR = Nu-SVR; PANAS = Positive and Negative Affect Schedule; PCA = patient classification accuracy; PHQ-9 = Patient Health Questionnaire-9; PPD = postpartum depression; PPV = positive predictive value; PR = pattern recognition; PSQ = Perceived Stress Questionnaire; PTSD = post-traumatic stress disorder; R&D = research and development; RCT = randomized controlled trial; RF = random forest; ROC = receiver operating characteristic; SCID = Structured Clinical Interview for DSM IV-TR; SCZ = schizophrenia; SLR = simple logistic regression; SSD = schizophrenic spectrum disorder; SVM = support vector machine; SVM-RFE = support vector machine learning with recursive feature elimination.

[a] One author is the founder of Koko (for-profit enterprise).

[b] Additional intervention was being asked, "Be honest, how likely are you to try the resources I just shared?". If "very likely" was answered, the patient continued application use as usual. If "not likely" was answered, an interactive barrier reduction intervention was presented (i.e., asking about potential barriers to use of crisis resources, then sharing information designed to help user overcome indicated barrier).

[c] The three-minute speech task was a standardized, adapted version of the Trier Social Stress Task for children seven and older. Participants were told to prepare a speech and were told they would be judged on how interesting the speech was. Only the three-minute speech was recorded on a video camera for use in the ML experiment.

Appendix 4: Critical Appraisal of Included Publications

Table 5: Strengths and Limitations of Systematic Reviews and Meta-Analyses using Amstar 2[16]

Strengths	Limitations
Burke (2019)[21]	
• The research questions and inclusion criteria for the review include some components of the population, intervention, and outcomes • Extensive databases searched with keywords provided • Population, type of ML, and results clearly described • Funding source of SR provided, conflicts of interest provided (none)	• The eligible control groups (i.e., healthy individuals, other mental disorders, etc.) were not specified in the methods • No a priori methods; i.e., no protocol or PROSPERO registration • No explanation for study design inclusion • No dates for search provided, no grey literature searched, no trial or study registries searched, unknown if search conducted within 24 months of publication • No information provided on data extraction or screening, unknown if performed by more than one reviewer • No list of excluded studies • Unknown what the comparators were for the studies (i.e., if sensitivity of ML was compared with psychiatric evaluation, *DSM-IV* criteria, chart notes, etc.) • No critical appraisal of studies, no risk of bias discussed, no heterogeneity discussed • No sources of funding provided
Colombo (2019)[20]	
• Eligible study designs provided, clear population and intervention • Search conducted within 24 months of publication • Protocol provided • Comprehensive search strategy with two or more databases • Risk of bias performed for EMI studies, by two independent reviewers, with specific tool provided • Three reviewers independently screened titles and abstracts and selected full-text papers • Study design inclusion explained • Included studies described in detail • Funding source of SR provided, conflicts of interest provided (none)	• Date of search provided, but date range of the search not provided; unknown if search was from database inception or later • Eligible comparators or outcomes not provided • No risk of bias assessment performed for EMA, despite justification for not performing assessment • No information on who performed data extraction (i.e., duplicate screening) • No justification for exclusion of non-English publications • No list of excluded studies • No grey literature searched, no trial or study registries searched • No sources of funding provided • Risk of bias or critical appraisal not discussed in results, no discussion on heterogeneity in studies

Strengths	Limitations
De Filippis (2019)[24]	
• Eligible populations, comparators, and interventions clear • Search conducted within 24 months of publication • Comprehensive search strategy with two or more databases, searched reference lists of included studies • Two reviewers independently screened titles and abstracts and selected full-text papers • Included studies described in detail • Risk of bias assessed using the Jadad scale • Only included studies with a Jadad score of 3 or greater, limiting potential risk of bias • Potential for high heterogeneity briefly discussed • Conflicts of interest provided (none)	• No explanation for study design inclusion • No justification for exclusion of non-English publications • No information on who performed data extraction • No list of excluded studies • No grey literature searched, no trial or study registries searched • No sources of funding provided • Risk of bias or critical appraisal not discussed in results • Funding source of SR not provided
Passos (2019)[23]	
• Search conducted within 24 months of publication • Comprehensive literature search of more than two databases, no language restrictions, searched reference list of publications • Inclusion criteria of population, intervention, outcomes included • Conflict of interest stated • Duplicate screening performed • Funding source of SR provided	• No list of excluded studies • No protocol provided • Unknown if grey literature searched • No risk of bias assessed • No limitations on study design discussed, no information on study designs of included studies • No funding sources of primary studies included • No discussion of bias in results
Gao (2018)[7]	
• Search conducted within 24 months of publication • Eligible population and intervention clear • Comprehensive literature search of more than two databases, searched reference list of publications • Common sources of bias in diagnostic studies discussed • Included studies described in detail • Funding source of SR provided, conflicts of interest provided (none)	• Abstract screening by one reviewer • Inclusion criteria not clear • No language or publication year restrictions, no grey literature searched • Sources of funding not provided • No list of excluded studies • Unknown if screening or data extraction done independently by two reviewers • No risk of bias assessed

Strengths	Limitations
Laranjo (2018)[13]	
• Search conducted within 24 months of publication • Eligible population and intervention clear • Comprehensive literature search of more than two databases, no language or publication year restrictions, grey literature searched, searched reference list of publications • Registered prior to review completion in PROSPERO • Full-text screening by two reviewers • Risk of bias assessed using Cochrane tool and discussed in results • Meta-analysis not attempted due to heterogeneity of outcomes and interventions • Data extraction reviewed by two independent reviewers • Included studies described in detail • Sources of funding provided • Funding source of SR provided, conflicts of interest provided (none) • List of excluded studies with reasons	• Abstract screening by one reviewer
Lee (2018)[22]	
• Search conducted within 24 months of publication • Comprehensive literature search of more than two databases, no language or publication year restrictions, grey literature searched, searched reference list of publications • Clear inclusion criteria of population, intervention, outcomes, and designs • Heterogeneity assessed with pre-specified cut-offs • Studies described in adequate detail • Publication bias assessed using Egger's and Begg's tests with the trim and fill method • Used random-effects model for meta-analysis • Heterogeneity sources and limitations discussed in conclusions • Funding source of SR provided, conflicts of interest provided (none)	• No specified eligible comparators (e.g., comparisons to healthy controls, other mental health disorders, other treatments, no treatment) • Only included diagnosed patients with depression, (excluded at-risk patients), which may limit generalizability • Limited by language (English only) • No list of excluded studies • No information on duplicate data extraction or screening • No protocol provided • Combined studies into meta-analysis with extremely high heterogeneity (92%, significant) with different predictors, different treatments, and different machine learning models (92% were supervised) that may not have been appropriate, did not justify the inclusion of these studies together • Risk of bias not discussed in results • No funding sources of primary studies included

Strengths	Limitations
Wongkoblap (2017)[25]	
• Search conducted within 24 months of publication • Comprehensive literature search of more than two databases, no language limitations, searched reference list of publications • Inclusion criteria of population, interventions included • Conflict of interest stated (none)	• No list of excluded studies • No information on duplicate data extraction • No protocol provided • No risk of bias assessed • No limitations on study design discussed, no information on study designs of included studies • No list of excluded studies • No information on duplicate extraction or screening • No funding sources of primary studies included • No discussion of bias in results • Not a lot of detail on study results provided

DSM-IV = Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition; EMA = ecological momentary assessment; EMI = ecological momentary interventions; ML = machine learning; SR = systematic review.

Table 6: Strengths and Limitations of Clinical Studies Using the Down's and Black Checklist and QUADAS2[17,18]

Strengths	Limitations
Jaroszewski (2019)[35]	
• The hypothesis, aim, and objective of the study were clearly described • The main outcomes were clearly described in the methods • The interventions of interest were clearly described (Koko app version 1 vs. version 2) • The main findings of the study were clearly described • The main outcome measures used were accurate • Reasons included in the BRI were created prior to the study through analysis of a random set of responses to identify common reasons for the avoidance of crisis services, so it was likely that the multiple choices categories provided to the participants were accurate • Decisions to include or not include barriers were made a priori and justified • Chi-square tests were used to analyse categorical outcomes • Missing data were tested to determine if missing at random or dependent on baseline characteristics • Actual probability values been reported for the main outcomes • The participants that took part in the study were likely representative of the entire population from which they were recruited as all Koko users were included in the study • The patients in different intervention groups were recruited from the same population • No blinding, but this was unlikely to affect the results of the study as each app was used individually, and there was no administration of interventions by staff	• No demographic information was collected from the actual users in the intervention as the intervention was anonymous; separate demographic information was surveyed in the same time frame as the intervention, so may represent the users of the platform during the time period, but the number surveyed was significantly smaller than the number of users (496 surveyed vs. 39,450 users); it was therefore unknown if the groups differed at baseline, but it was unlikely as the groups were randomly generated • No confounders taken into account as no confounding variables were measured prior to the intervention • No adverse events reported, but unlikely to be possible in the study design • A large number of participants did not respond to the follow-up assessments (more than 50% in both groups), potentially leading to attrition bias • Power was limited in some subgroup analyses due to smaller sample sizes (e.g., in the outcome of crisis resource usage, one subgroup was individuals who indicated they did not plan to use crisis resources who then looked at one of the psychoeducation resources provided by the application) • Conclusion limited to individuals who were aware of the Koko app, or to individuals who were more likely to use technology-based applications • Misclassification may have occurred with the "in crisis/not in crisis" group as the app did not use objective metrics (e.g., suicide attempt) to estimate a probability of crisis • Outcome measurements (i.e., "did you use crisis resources?") may be subject to bias through social desirability and self-report • It was unknown if individual who used crisis resources were in an area that has greater availability of these resources • Individuals in the intervention group received the question "are you likely to use crisis resources," which potentially primed this group to think about or to use crisis resources. Individuals in the control group did not receive this question • Not all individuals in the intervention groups received the exact same intervention; i.e., an individual who answered "not likely" to use crisis resources received the BRI and those who answered "likely" did not receive the BRI

Strengths	Limitations
Breen (2019)[27]	
• Control groups included trauma survivors without PTSD to ensure the classification was based on PTSD and not the presence of trauma • Power calculation performed • Order of sleep sessions was randomized • Reference standard was a validated PTSD scale and neuropsychotic inventory • Validation with LOO cross validation, and compared with other ML methods • Groups were controlled for age, nicotine use, and HIV status	• Unknown how much time passed between reference test and index test; the sleep test results may differ from the original questionnaires provided • Demographics recorded but no detail on whether groups differed from one another in other, unrelated variables (i.e., not in the presence of trauma or PTSD, or related factors) • Thresholds for the models were not clear
Byun (2019a)[28]	
• All experimental procedures were performed in the same humidity and temperature-controlled room • Validation with LOO cross validation • Reference standard the current gold standard for diagnosis — psychiatrist evaluation • Groups were age and gender matched • Detail was provided on the demographics of groups and statistical comparison • Thresholds for the models were clear	• Unknown how much time passed between reference test and index test; the sleep test results may differ from the original questionnaires provided • Patients with depression were on antidepressant medication, which may affect heart rate variability, while patients in the control group were not
Byun (2019b)[29]	
• LOO method used to validate data • Demographic information provided for control and MDD groups, with statistical testing (no differences except HAMD score, which was expected) • Reference standard was the HAMD and the *DSM-IV* criteria evaluated by a psychiatrist • Participants were matched for age and gender • All subjects underwent the same procedure for measurements • Conflicts of interest stated (none)	• Small number of participants in study (n = 66); with a heterogenous population such as patient with depression, this may have limited the reliability of results • Patients with depression were on antidepressant medication, which may affect heart rate variability, while patients in the control group were not • Unknown if sampling was done by convenience, or what reason the controls were attending the medical centre; also, unknown if patients attending the medical centre for reasons related to their depression or for unrelated reasons • Diagnostic accuracy may be overinflated as the enrolled patients had a known disease prior to enrolment; the control groups were also known to not have depression prior to enrolment

Strengths	Limitations
Deng (2019)[30]	
• Objective of study was clear • SCZ was diagnosed with structural clinical interview by psychiatrists • Participants were matched for age and gender • All subjects underwent the same MRI procedure and same screening tests • LOO cross validation performed • Demographics provided	• Diagnostic accuracy may be overinflated as the enrolled patients had a known disease prior to enrolment; the control groups were also known to not have SCZ prior to enrolment • Patients with schizophrenia were given SCZ-related medication after assessment but prior to the imaging, which may have affected the imaging scans • The groups significantly differed in educational levels and in IQ scores. IQ score differences may be linked with SCZ; therefore, this was a potential confounding item that was not addressed
Ding (2019)[32]	
• Demographics provided with statistical testing between groups • Individuals taking medication (clozapine or chlorpromazine) that affect brain function were excluded • MDD was diagnosed through clinical interviews by psychiatrists • Participants were matched for age and gender	• Patients in the control group were recruited from the community, while MDD groups were recruited from the outpatients of a hospital • Unknown how the 80% training and 20% testing amount of sample was decided on • Diagnostic accuracy may be overinflated as the enrolled patients had a known disease prior to enrolment; the control groups were also known to not have depression prior to enrolment
Kalmady (2019)[37]	
• 5 times 10-fold cross validation performed • SCZ was diagnosed with structural clinical interview by psychiatrists • Controls were recruited from the same location, with the same screening tests for the patients with SCZ • Patients had never been treated with psychotropic medications, so control patients and patients with SCZ had similar medication backgrounds • Participants were matched for age and gender	• Diagnostic accuracy may be overinflated as the enrolled patients had a known disease prior to enrolment; the control groups were also known to not have SCZ prior to enrolment • No demographics were provided
Leightley (2019)[38]	
• PTSD was diagnosed using the PTSD Checklist Civilian Version based on the *DSM* • 10-fold cross validation performed • Demographics provided with no statistical testing	• Data set used was not previously created for use in building an ML model • Questionnaire for determining PTSD performed at a different time than the data collection for the ML model • Unknown what the thresholds were for the models

Strengths	Limitations
McGinnis (2019)[40]	
• "Low quality" data used as the testing data set for one accuracy test; This was independent from the "high quality" data set which was used for training the model • Consecutive sample of patients enrolled • Sample was not gathered based on previously confirmed diagnoses; i.e., children were not selected because they were previously diagnosed with internalizing disorders or previously deemed a "healthy" control • Patients with suspected autism spectrum disorders (a condition which can affect speech) were excluded from the study • Parameters for the speech (feature selection) were based on previous tests in adults, not arbitrarily chosen • Algorithm was validated with the LOO cross-validation technique • Chosen reference standard the current gold standard and validated method for classifying the groups • Unknown how long of an interval between the clinical interview and the speech task, but unlikely to have affected the results (e.g., no spontaneous change in condition status) • No loss to follow-up	• The outcomes for accuracy in the "high quality" data may not be truly accurate, as the algorithm was trained on that data set; the poorer performance in the "low quality" data set is not known to be due to the low quality in the data, or if the high-quality trained model was overfit to that data set, and therefore does not generalize well to other datasets • The three-minute speech task was occasionally conducted with siblings in the room with the participant, which introduced variability into the conduct of the task across participants • The standard test requires the "audience" to remain disinterested and critical of the speech task; the introduction of factors such as siblings and caregivers may have prevented the task from being performed to the intended standard; the methodology stated the test was "standardized," but was not performed in the same room, with the same experimenters, or the same conditions • The three-minute speech task was shown to induce anxiety in children seven and older; the children in the present study were aged three to seven; it was unknown if the task was suitable for children of this age group • The data that were trained for the algorithms were classified by one researcher
Mohr (2019)[41]	
• The hypothesis, aim, and objective of the study were clearly described • The main outcomes were clearly described in the methods • The characteristics of the patients included in the study were clearly described • The interventions of interest were clearly described (coached vs. self-guided and weekly app recommendations vs. no recommendations) • The main findings of the study were clearly described • Actual probability values been reported for the main outcomes • The staff, places, and facilities where the patients were treated were likely representative of the treatment the majority of patients receive • Due to the nature of this study, patients were not blinded to their designated treatment, which was appropriate • The statistical tests used to assess the main outcomes were appropriate • The main outcome measures used were accurate • Study subjects were randomized to intervention groups • Study subjects in different intervention groups were recruited over the same period of time • Losses of patients to follow-up were taken into account, although the amount of patients lost to follow-up was low • The study had sufficient power to detect a clinically important effect where the probability value for a difference being due to chance was less than 5%; a power calculation was performed	• No confounders taken into account as no confounding variables were measured prior to the intervention • No adverse events reported, but unlikely to be possible in the study design • The patients that participated in the study were not likely representative of the entire population from which they were recruited as subjects were only selected if they had an Android phone and were excluded if they used another type of phone (e.g., iPhone) as it may not be compatible with the app; this may limit the generalizability of the study in the overall population; additionally, the rigorous screening and consulting procedure for subject participation may a created a sample that was more motivated to engage in digital mental health than the regular population • There were some issues within the coaching groups as some subjects did not receive proper treatment as a result of coaches not following through with intervention • No follow-up data were collected in the study • Potential confounders were not stated by the authors and how they may affect the results or interpretation of the study

Strengths	Limitations
Oh (2019a)[42]	
• Provided a clear definition of "machine learning" and "deep-learning" • Study aim was clearly identified in the text • Described the two data sets where the data were obtained from • Data from the NHANES data set included longitudinal information collected from 15 years of data collection • Both NHANES data set and K-NHANES data set provided a large sample size • Development and validation of ML and DL algorithms were clearly described • 10-fold cross validation was used to develop ROC curves • DL classifier was measured against more than one ML classifier • Limitations of the study were discussed	• Data from the K-NHANES data set only included one year of data, which was not as robust compared with the NHANES data set • Both NHANES and K-NHANES largely used self-reported data • Characteristics and demographics of the data used were not identified • Depression was measured in a binary manner so severity of depression cannot be interpreted • Cross-sectional results differed between NHANES and K-NHANES, which may be due to cross-national diversity or cultural differences
Oh (2019b)[43]	
• Study aim and objectives were clearly indicated in the text • SSD participants were evaluated on *DSM-IV-TR* criteria and inclusion and exclusion criteria was outlined • HC participant recruitment process and inclusion and exclusion criteria were identified • Subject consent and study approval was addressed • All participants performed the same experimental task • Data processing, ML model, and classifier evaluation approaches were all described in text and with visual frameworks • Clinical implications of classification performance were highlighted • Study limitations were addressed • Potential modes of bias were mentioned	• Recruitment location for SSD participants was not identified • There was a much larger number of SSD participants compared with HC participants • The number of subjects in training and test phase was small • Because of limited sample size, cross validation was not possible for providing an almost unbiased estimate of error • Authors did not provide final study conclusion
Ramkiran (2019)[44]	
• Study objectives were clearly outlined in the text • Initial data set screening was done and a sample population with an equal number of SCZ and matched HC participants was yielded • Participant inclusion process was outlined • Data acquisition and data processing were clearly described and justified • SVM model application and use was described • Classification outcomes and clinical implications were addressed • Results were similar with previous studies • Future study considerations were presented in the discussion • Confounding for demographics and clinical variables were mentioned • Study limitations were addressed and possible solutions were mentioned • Authors clearly stated the study conclusion	• Small number of samples available for both training and testing set, which may have impacted classificatory capacity • Access to complete demographics for the data set was not available • Analysis was carried out on a single data set, which does not express heterogeneity

Strengths	Limitations
Schwarz (2019)[45]	
• Study aims were clearly presented in the text • A leave-site-out procedure was used to create a training and test data set • Large study population with eight different cohorts • Demographic and recruitment details were provided in supplementary information • Participant consent and study approval was provided • Classification data were analyzed using RF and SVM	• Study hypothesis was not clearly outlined • Data analysis did not factor in antipsychotic medication, which has been known to alter brain matter • Study limitations were not clearly outlined
Walsh-Messinger (2019)[46]	
• Study objectives and hypothesis were clearly identified in the text • Study population was adapted from a longitudinal study funded by NIMH studies • Inclusion criteria for HC participants was outlines • Approval and participant consent were identified • Participants characteristics and assessment results were outlined • All participants underwent the same comprehensive assessments • Initial ML model screening was performed to decide which model was most appropriate to use for the analysis • HC misclassification was addressed • Results of classification analyses were clearly described and presented in the text and performances were appropriately measured • Study limitations were addressed and potential solutions were presented	• Sample technique for sample population was not discussed • Inclusion and exclusion criteria for psychiatric participants were not outlined • Retrieval of clinical research data of sample population was not mentioned • More data were collected from one data site (site I) versus the other (site II) • Case and controls groups had a significant difference for age and diagnostic distribution, which may have influenced results
Wang (2019)[48]	
• Background information was clearly presented and used for the study justification • The purpose and goal of the study was clearly identified in the text • The study process was outlined in the introduction • Data obtained for the study was outlined and data parameters were described (retrieval location, dates, justification, and data characteristics) • Study population was described and representative of the target population • Study population exclusion criteria was clearly outlined • Baseline characteristics of the study population was clearly displayed and described • Study outcomes and predictors were fully described • Prediction ML models were clearly identified and described • Results of the prediction models were clearly presented and described • Performance was appropriately measured using AUC, sensitivity, and specificity • Study outcomes were highlighted and discussed • Study outcomes were consistent with previous studies	• Electronic health records used as data were collected from a single health system • Study population did not control for patients on medication (antidepressants), which may introduce bias into the results • Not all pregnant patients underwent assessment post-birth. It was assumed that all patients who did not receive a diagnosis with the ICD-10 codes O99.3 and O99.34 did not have PPD. This may not be accurate as not all patients are treated for PPD or receive an appropriate diagnosis.

Strengths	Limitations
Zhao (2019)[49]	
• Study hypothesis was clearly outlined • Participant consent was obtained • Participants completed a GAD-7 and PHQ-9 questionnaire • All participants had their walking gait recorded • Five different ML algorithms were used and a 10-fold cross validation was applied • ML were both linear and nonlinear regressed • Limitations for this study were addressed • Authors discuss appropriateness for future studies	• Study aim and approach were not clear • Background evidence to support the study and study feasibility was questionable • Sampling technique was not clearly stated • Sample population was graduate students, which was not representative of a clinical population • Number of male and female participants was not equal • Only 167 out of 179 PHQ-9 scores were used and there was no explanation as to why • Questionnaire scores were not validated by a clinician • A small number of participants had a questionnaire score that would be considered severe anxiety or depression, so validity of classification models was questionable • Training and testing models were applied to males and females separately, which may have impacted the results found • Results to support the main outcome (evaluating an individuals walking gait to predict anxiety and depression levels) was not clear • Results were not generalizable due to the sample population (graduate students) • Solutions to study limitations were not clearly identified • Potential confounders and bias were not discussed
Zhuang (2019)[51]	
• The aim, process, and hypothesis of the study was clearly outlined in the text • FES and HC participants were matched for age, gender, and education level • Recruitment location and screening process were stated • Exclusion criteria was outlined • FES patients had no comorbidities • Study approval and participants consent was described in the text • Data acquisition, data processing, feature selection, classification models, and performance evaluation were all outlined and described • Process for classification sequence was outlined using a flow diagram framework • Results for all outcomes were shown using figures and tables • Study limitation were discussed and future considerations were mentioned • Study conclusion was clearly stated	• Small sample size • Unequal number of participants in case and control arms • Participants recruitment technique was not clearly described • Was not clear if FES and HC participants were recruited from the same location • “Drug naïve FES” was not clearly defined • Possible confounders or risk of bias was not addressed in the text

Strengths	Limitations
Fulmer (2018)[33]	
• The hypothesis, aim, and objective of the study were clearly described • The main outcomes were clearly described in the methods • The characteristics of the patients included in the study were clearly described • The interventions of interest were clearly described (two weeks of Tess, four weeks of Tess, or information-only control group) • The main findings of the study were clearly described • Actual probability values were reported for the main outcomes • The staff, places, and facilities where the patients were treated were likely representative of the treatment the majority of patients receive • Due to the nature of this study, patients were not blinded to their designated treatment, which was appropriate • The statistical tests used to assess the main outcomes were appropriate • The main outcome measures used were accurate • Study subjects were randomized to the intervention and control groups • Study subjects in different intervention groups were recruited over the same period of time • Losses of patients to follow-up were taken into account, although the amount of patients lost to follow-up was low • The study had sufficient power to detect a clinically important effect where the probability value for a difference being due to chance was less than 5%; a power calculation was performed	• The subjects that participated in the study were not likely representative of the entire population from which they were recruited; however, two-thirds of the subjects were women and the majority of the subject where Asian or white and may not be representative of the general population • The authors acknowledged that the control group experienced an increase in anxiety symptoms, which may be due to confounders that were adjustable • Did not report all results, did not provide specific numbers for PHQ-9, GAD-7, or PANAS after intervention
McGinnis (2018)[39]	
• Consecutive sample of patients enrolled • Sample was not gathered based on previously confirmed diagnoses; i.e., children were not selected because they were previously diagnosed with internalizing disorders or previously deemed a “healthy” control • Algorithm was validated with the leave-one-out cross-validation technique • Chosen reference standard the current gold standard and validated method for classifying the groups • No loss to follow-up	• Only accuracy was provided as an outcome, and AUC only plotted for two models • Thresholds used for the accuracy measurements were unclear • Two patients excluded due to technical difficulties, but what these issues were was not detailed

Strengths	Limitations
He (2017)[34]	
• The purpose, process, and objectives of this study were clearly presented in the text • Large sample size (N = 300) with equal number of participants in the case (PTSD) and control (non-PTSD) arms • PTSD participants were identified and diagnosed by two clinical practitioners using standardized instruments • Age and gender characteristics were reported for the sample population • Textual data processing included a preparation, training, and test phase • Instruments and methods used for textual data processing were clearly described • Reasoning and explanations for the use of each ML model was clearly outlined • 10-fold cross validation was used and performance was measured using accuracy, SE, SPE, PPV, and NPV • Performance measurements were clearly defined relevant to the study • Comparator for classifiers was outlined and explained • Results were clearly presented within the text and supplementary figures were provided • Two main limitations were identified and potential solutions were outlined	• Sampling technique for study participants was not outlined • Did not state whether case and controls were matched in any way • Data collection was based on an online forum, which could introduce bias (no real way of knowing if the participants personally wrote forum responses) • This study favoured individuals with access to a computer • As the sample was representative of people seeking mental health care, this may not be generalizable to the PTSD population • As this study looked at natural language processing, only individuals with English as their primary language were involved
Wang (2017)[47]	
• Reference standard for the model was actual measured data as the study was retrospectively analyzed, so the reference standard was reliable for that time period • Cost accuracy in addition to patient accuracy used as an outcome as it can show the relative cost for patients • Explanation for cut-offs of 20% and 10% (10% of patients use 50% of resources; 20% used 70%) • Model trained on separate training data set • Conflicts of interest declared • 10-fold cross validation used	• Conclusions limited for groups of patients on other insurance plans (e.g., Medicaid), and individuals not yet diagnosed with SCZ • Authors employees of the funding body • No residual plots provided, so unknown if predicted values were biased

Strengths	Limitations
Devine (2016)[31]	
• Longitudinal design allows for comparison of reliability over time, as opposed to cross-sectionally in time • Population studied likely representative of those who would receive the CAT in a real-life setting • CATs were developed separately from this application, so had been trained and validated in a separate population than the tested population • Static assessments used were validated tools for measuring depressive symptoms, stress, and anxiety, and were therefore appropriate choices for comparison	• CAT has physical symptom questions removed, creating a unidimensional item bank • There was no control group in the study; all patients had a diagnosis and were in an in-patient setting, so it was unknown the accuracy of the CAT measure (only precision was measured) • No sample size calculation • Unknown what the threshold was
Achtyes (2015)[10]	
• Population with mental health disorders in the study likely representative of those who would receive the CAT in a real-life setting • CATs were developed separately from this application, so had been trained in a separate population than the tested population • Threshold for diagnosis set at 50% • Comparison assessments used were validated tools for measuring DSM conditions, and therefore were appropriate choices for comparison	• Sample size calculation (n = 150 minimum), but only 145 patients enrolled in the study (did not account for potential dropout or withdrawn consent) • When restricting the groups to test accuracy (for example, limiting groups to only samples of patients with depression) sample size was very small relative to the required sample size calculation • Patients with SCZ or psychotic disorder excluded, but reasoning not explained (maybe because of the types of CAT tests used) • Participants were a convenience sample from the website and clinic waiting rooms, and therefore only included patients who were aware of the Pine Rest outpatient clinic; therefore, it was unknown whether the control groups were representative of a control group that would occur in a real-life setting; additionally, patients receiving treatment may be considered more “severe” in their mental illnesses, which may make the test appear more accurate than it would be if given to a random sample of people • CAT-MANIA had never been validated in a clinical sample prior to this study • Only 19 healthy controls included, in comparison to relatively more patients with mental health conditions • Unknown if CAT tests differentiated between different mental health diagnoses, only accuracy with respect to the healthy controls reported; this design may overinflate the accuracy of the tests • Unknown how long delay between two tests was, or if there was variation in the order in which they were presented (i.e., to prevent participant fatigue, or prevent participants from learning the types of questions on one test)

Strengths	Limitations
Jimenez-Serrano (2015)[36]	
• Model was trained and validated on a different data set than the testing data set • Tested models other than the previous study's test had identical numbers of variables and weeks • Population used likely to represent the general population that would be at risk for postpartum depression, and not likely to only include patients with severe illness • Tools used for comparison were validated tools used in detecting postpartum depression • Unknown how much time passed between reference test and index test, but not relevant in this design of test	• Unclear exactly what the reference standard was, appears as though some individuals got an EPDS and a DIGS test while some (who were negative on the EPDS) did not receive the DIGS test; the reference standard should be equal across all groups • Threshold for the models were not clear for all the models and tests; only the "best" thresholds shown, which may overoptimize the models
Zhou (2015)[50]	
• 10-fold cross validation used • Model was trained and validated on a different data set than the testing data set • Patients were randomly selected from hospitalization data; sample was not gathered based on previously confirmed diagnoses of depression • No loss to follow-up due to study design	• Only patients with heart conditions used, so may not generalize to all patients with depression • Reference standard not a validated tool nor a mental health specialist (a pharmacy student and an internal medicine physician), so unknown whether mistakes due to human error occurred • No demographics provided

AUC = area under the curve; BRI = barrier reduction intervention; CAT = computerized adaptive test; CAT-MANIA = computerized adaptive test for manic/hypomanic symptom severity; DIGS = diagnostic interview for genetic studies; DL = deep learning; *DSM = Diagnostics and Statistics Manual of Mental Disorders; DSM-IV = Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition; DSM-IV-TR = Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition, Text Revision*; EPDS = Edinburgh Postnatal Depression Scale; FES = first-episode schizophrenia; GAD-7 = Generalized Anxiety Disorder 7-item scale; HAMD = Hamilton depression rating score; HC = healthy controls; ICD-10 = International Statistical Classification of Diseases and Related Health Problems 10th Revision; IQ = intelligence quotient; K-NHANES = Korea–National Health and Nutrition Examination Survey; LOO = leave one out; MDD = major depressive disorder; ML = machine learning; MRI = magnetic resonance imaging; NHANES = National Health and Nutrition Examination Survey; NIMH = National Institute of Mental Health; NPV = negative predictive value; PANAS = Positive and Negative Affect Schedule; PHQ-9 = Patient Health Questionnaire-9; PPD = postpartum depression; PPV = positive predictive value; PTSD = post-traumatic stress disorder; RF = random forest; ROC = receiver operating characteristic; SCZ = schizophrenia; SE = sensitivity; SPE = specificity; SSD = schizophrenic spectrum disorder; SVM = support vector machine; vs. = versus.

Appendix 5: Main Study Findings and Authors' Conclusion

Table 7: Summary of Findings of Included Systematic Reviews and Meta-Analyses

Main Study Findings	Authors' Conclusion
Burke (2019)[21]	
N = 35 studies **Outcome: Suicide death** **N = 5 primary studies** Three studies for suicide death among patients with varying psychiatric histories Service members with baseline psychiatric hospitalization: Regression trees, elastic net regression, LASSO, naive Bayes, RF, support vector regression, elastic net penalized regression (comparator NR) AUC = 0.71 to 0.89 Service members with baseline outpatient mental health visit, 26 weeks after visit: Naive Bayes, RF, support vector regression, elastic net penalized regression (comparator NR) AUC = 0.72 (patients with prior psychiatric hospitalizations) AUC = 0.61 (patients without hospitalizations) AUC = 0.66 (combined sample) Veterans who died by suicide (identified at the top 0.1% and 1% of suicide risk) and time-matched controls: Elastic net, DT (Bayesian additive regression trees, RF), spline (adaptive splines, adaptive polynomial splines), generalized boosted regression models (adaptive boosting), SVM (linear kernel, polynomial kernel, radial kernel) (comparator NR) Sensitivity = 2.7% Specificity = 10.7%	"The use of ML techniques for improved prediction may be most notable in predicting the outcome of suicide death as, despite its importance, limited research has focused on this low base rate outcome. Among the included reviewed studies, AUCs ranging from 0.71 to 0.89 were achieved in predicting suicide death… Importantly, improved model accuracy with the employment of ML for the prediction of suicidal behavior has been achieved in prediction windows as short as 7 days and as long as 2 years, demonstrating its potential use in informing both crisis intervention and long-term prevention" (p. 880).

Main Study Findings	Authors' Conclusion
Patients at high risk for suicide: Bayesian Dirichlet equivalent DT (comparator NR) Individuals who identified as non–African-American with a substance use disorder and who were psychiatrically hospitalized within the past year were at highest risk for suicide NLP in unstructured clinical notes for veterans: Supervised training with genetic programming; unspecified machine learning algorithm (unspecified) (comparator NR) Accuracy (single words) = 46% to 65% Accuracy (single words and phrases) = 52% to 69% **Outcome – SA** N = 14 primary studies One study examined adolescents, 13 studies adults (various characteristics) One longitudinal study, 13 cross-sectional Adults who had an EMR documented self-injury code (ML models based on EMR indicators): Random forests (comparator NR) AUC = 0.80 to 0.84, model performance increasing closer to time of SA (720 days to 7 days prior to SA) Adults who were either hospitalized or admitted to the emergency department due to suicidal behaviour: Differentiating between adults with first SA and no SA (DT vs. NR): Accuracy = 81.4% Sensitivity = 0.87 Specificity = 0.86 Precision = 0.86	

Main Study Findings	Authors' Conclusion
SA versus controls (including participants reporting ideation): RF, naive Bayes, SVM, predictive association rules, DT, neural networks (comparator NR) Sensitivity = 0.95 to 0.96 PPV of 0.93 to 0.97 Biological factors that may predict lifetime SA among males 3 single-nucleotide polymorphisms had most powerful explanatory power in SA Forward selection, SVM (comparator NR) Sensitivity = 0.54 Specificity = 0.80 Positive likelihood ratio = 2.71 Negative likelihood ratio = 1.75 Sample diagnosed with mood, SCZ spectrum, or personality disorders Predict recent and remote SA history, ideation was the most important predictor of recent SA status (DT [comparator NR]) (AUC = 0.80, sensitivity = 0.73, specificity = 0.80, PPV = 0.58) Lifetime aggression was the most important predictor of remote SA status (AUC = 0.65, sensitivity = 0.89, specificity = 0.36, PPV = 0.44) Samples receiving outpatient mental health care MDD or bipolar, lifetime SA history LASSO, SVM, relevance vector machine (comparator NR) Accuracy = 65% to 72% AUC = 0.77 Sensitivity = 0.72 Specificity = 0.71	

Main Study Findings	Authors' Conclusion
SCZ spectrum disorders, lifetime SA history LASSO, RF, support vector classifier, elastic net (comparator NR) AUC = 0.71 Accuracy = 0.67 Sensitivity = 0.64 Specificity = 0.68 Depression and Anxiety disorders Artificial neural network (comparator NR) One-month SAs (AUC = 0.93, accuracy = 93.7%, sensitivity = 0.12, specificity = 0.99) Past one-year model (AUC = 0.89, accuracy = 90.8%, sensitivity = 0.33, specificity = 0.98) Lifetime model (AUC = 0.87, accuracy = 87.4%, sensitivity = 0.77, specificity = 0.91) Community samples Adolescents, past year SA RF (comparator NR) Past year SA = 90% accuracy Asian-Americans, lifetime SA DT (comparator NR) Sensitivity = 0.75 Specificity = 0.39 PPV = 0.39 NPV = 0.75 Undergraduates with history of NSSI, lifetime SA Elastic net regression, DT, RF (comparator NR) AUC = 0.75	

Main Study Findings	Authors' Conclusion
Outcome: Suicide planning N = 1 primary study Recent suicidal planning among those with a history of NSSI Elastic net regression, DT, RF (comparator NR) AUC = 0.89 **Outcome: Suicidal Ideation** N = 10 primary studies 4 longitudinal studies, 6 cross-sectional studies Adolescent SI at one-year follow-up DT (comparator NR) Sensitivity = 0.47 to 0.78 Specificity = 0.68 to 0.91 NLP in adults after hospital discharge for a suicide-related event NLP (comparator NR) Structured indicator accuracy: Sensitivity = 0.76 Specificity = 0.62 PPV = 0.73 Unstructured indicators accuracy: NLP (comparator NR) Sensitivity = 0.56 Specificity = 0.57 PPV = 0.61	

Main Study Findings	Authors' Conclusion
Adult sample All models, past two-week SI: AUC > 0.80 Simplest model (model NR, comparator NR), past two-week SI: Overall AUC = 85.6 Sensitivity = 0.78 Specificity = 0.83 PPV = 0.39 NPV = 0.97 Veterans, gender differences in SI DT, RF (comparator NR) Male: AUC = 0.91 Females: AUC = 0.92 Undergraduate students with a NSSI history Elastic net regression, DT, RF (comparator NR) AUC = 0.85 Use of fMRI to classify current SI among undergraduate students Multivoxel analysis (comparator NR) Accuracy = 91% Sensitivity = 0.88 Specificity = 0.94 PPV = 0.94 NPV = 0.89	

Main Study Findings	Authors' Conclusion
Outcome: Suicide risk N = 8 primary studies Mental health patients diagnosed with mood disorders AdaBoost, DT, knn, RF, neural network multi-layer perceptron, SVM (comparator NR) Accuracy = 78% Sensitivity = 0.77 Specificity = 0.78) N = 3, passive and active data collection methods for prediction through social media Social media posts (NLP), classifying patients based on "high" suicide risk RF (comparator NR) Recall value = 0.82 Note: recall value was the number of true positives and total number of positive instances Using Twitter (NLP) to predict suicide risk DT (comparator NR) Sensitivity = 0.53 Specificity = 0.97 PPV = 0.75 NPV = 0.93 Linguistic features of social media among Chinese adults SVM (comparator NR) "Poor ML performance in predicting suicide risk" p. 879 Among subset of users who had told others via social media that they wanted to kill themselves in the past 12 months: AUC = 0.61 Sensitivity = 0.65 Specificity = 0.58	

Main Study Findings	Authors' Conclusion
Questionnaires of psychological constructs among outpatient mental health patients to predict suicide risk DT (comparator NR) AUC = 0.59 Accuracy = 0.71 Precision = 0.73 Recall value = 0.63 Specificity = 0.79 Adolescent linguistic responses to open-ended questions and associated vocal characteristics; classifying the likelihood of presenting to the emergency room for SI or SA Cosine SVM (comparator NR) Accuracy = 96.67% SITBs vs. psychiatric controls, linguistic and vocal responses to differentiate between the two groups SVM (comparator NR) AUC = 0.82 **Outcome: Non-suicidal self-injury** N = 3 primary studies Undergraduate students Two ML techniques for identifying important indicators of NSSI frequency: Lasso regression, RF (comparator NR) R2 = 0.48 and 0.46 **Direct comparison to traditional methods** RF ML algorithm AUC (0.80 to 0.84) vs. multiple logistic regression AUC (0.66 to 0.68)	

Main Study Findings	Authors' Conclusion
Colombo (2019)[20]	
N = 1 primary study Mobylize! machine learning EMI (type of ML not specified, no comparator [single-arm trial]) n = 7 patients (7 out of 8 participants completed the 8-week program) • Mobylize! significantly reduced depressive symptoms (no data provided, *P* = NR) on PHQ-9 and clinician-based evaluations (QUIDS-C) • Anxiety symptoms were reduced (no data provided, *P* = NR) on GAD-7 • Predictive models did not reach high accuracy for mood (accuracy = NR) • Predictive accuracy for location, conversational state, and social interaction between 60% to 90% • Satisfaction with application rated 5.71 out of 7 (scale 1 to 7, 7 being high satisfaction) • 86% of participants noted that the intervention was helpful for identifying triggers and avoiding distressing behaviours	

Main Study Findings	Authors' Conclusion
De Filippis (2019)[24]	
N = 40 primary studies Studies looked at a variety of brain regions and networks and used fMRI 1.5 T to 3 T, rs-fMRI 1.5 T to 3.0 T, fMRI short/long-range FCs, and sMRI 1.5 T to 3 T sMRI studies • Accuracy ranged from 63.9% (SVM) to 88.4% (SVM-RFE) • AUC NR • Sensitivity ranged from 57.9% (SVM) to 91.9% (SVM-RFE) • Specificity ranged from 70.0% (SVM) to 87.0% (SVM) fMRI studies Accuracy ranged from 41% (MVPA) to 99.3% (ELM) AUC ranged from 0.61 (SVM) to 0.79 (SVM) Sensitivity ranged from 58.5% (DGM) to 100% (variety of ML models, MVPA, SVM, ROC, LIBSVM, VMHC) Specificity ranged from 40.9% (variety of ML models MVPA, SVM, ROC, LIBSVM) to 96.8% (variety of ML models, SVM, FC analysis, LIBSVM, ROC) *Note: In results with a variety of ML models, the specific model that produced the reported result was not specified*	"ML techniques represent a promising approach that could support clinicians in the diagnosis of mental disorders and may be useful in classifying SCZ through MRI. All studies included in the review achieved a minimum accuracy of approximately 60%, most of them between 75% and 90%, with differences between sMRI and fMRI, in favor of the second one (with accuracy peaks above 90%)" (p. 1624). "In conclusion, the application of ML techniques will be useful to automatically classify patients with SCZ on the basis of structural and functional MRI. If systematically included in the diagnostic process of patients with SCZ, these techniques could help physicians to detect patients, even in the early stage of the disorder, conferring an important clinical advantage. We imagine that the greater accuracy demonstrated by the various predictive models illustrated in this systematic review and new models resulting from the integration of multiple ML techniques will be increasingly decisive in the future for the early diagnosis and evaluation of the treatment response and to establish the prognosis of patients with SCZ" (p. 1624).

Main Study Findings	Authors' Conclusion
Passos (2019)[23]	
N = 90 primary studies Diagnostic studies using machine learning techniques in BPD Classification studies using structural neuroimaging and DTI SVM ML model Highest accuracies: 100% (BD vs. HC, BD vs. AD): 100% SEN and SPE (BD vs. HC, BD vs. AD) Lowest accuracy: 57% (BD); 54.76% (BD-I vs. psychotic MDD); 55% (BD vs. HC) SVM with RFE accuracy: 59.45% (MDD vs. BD) LICA ML model Accuracy: 66% (BD vs. HC); 58% (BD vs. SCZ) AUC: 067 (BD vs. HC); 0.59 (BD vs. SCZ) GPC ML model Accuracy: 69% to 78% (BD type I vs. HC); 65.5% to 79.3% (UD vs. BP) SEN: 64% to 77% SPE: 69% to 99% Lowest accuracy: 65.6% (unaffected HR vs. HC) RVM ML model Accuracies: 70.3% (WM); 64.9% (GM); 64% (WM + GM) SEN: 66.4% (WM); 58.6% (GM); 59 (WM + GM) SPE: 74.2% (WM); 71.1% (GM); 70% (WM + GM) AUC: 0.72 (WM); 0.7 (GM)	"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. 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" (p. 9).

Main Study Findings	Authors' Conclusion
Classification studies using functional neuroimaging SVM ML model Highest accuracy: 93.1% (BD vs. MDD); 81% (BD vs. HC) Best SEN: 92.0% (BD vs. MDD) Best SPE: 94.4% (BD vs. MDD) Lowest accuracy: 61.7% (BD vs. SCZ vs. HC) SVM-FoBa accuracy: 92.07% (BD vs. MDD); 82.22% (BD vs. HC) SVM-FoBa SEN: 86.36% (BD vs. MDD); 81.82% (BD vs. HC) SVM-FoBa SPE: 89.29% (BD vs. MDD); 82.61% (BD vs. HC) Multiclass SVM: 58% (HC); 64.5% (AR); 70.5% (BD) GPC ML model Highest accuracy: 83.5% (BD vs. unrelated HC) SEN: 84.6% (BD vs. unrelated HC) SPE: 92.3% (BD vs. unrelated HC) Lowest accuracy: 70% (BD vs. HC); 61% (BD: intense happy vs. neutral faces) rLDA ML model BD classification accuracy: 91.74% (2-back); 69.13% (0-back); 63.93% (GNG); 52.13% (FacePos); 49.55% (FaceNeg) Classification studies using genetic analysis RF ML model Highest accuracy: 0.734 (BD vs. HC) Lowest accuracy: 53.6% (BD vs. HC) SEN: 0.998	

Main Study Findings	Authors' Conclusion
NB ML model Accuracy: 0.702 (BD vs. HC) SEN: 0.734 KNN ML model Accuracy: 0.733 (BD vs. HC) SEN: 0.954 MDR ML model Accuracy: 0.647 (two-way MDR); 0.721 (three-way MDR) SEN: 0.664 (two-way MDR); 0.883 (three-way MDR) CART ML model Accuracy: 61% (HC vs. non-HC) SEN: 71% SPE: 50% CNN ML model Accuracy: 64.2% (BD vs. HC) DT ML model Accuracy: 54.8% (BD vs. HC)	

Main Study Findings	Authors' Conclusion
Classification studies using electroencephalographic measures SVM ML model Accuracy (BD in depressive episode vs. UD): 62.37%; 73.26% (PSO-SVM); 75.24% (GA-SVM); 78.21% (ACO-SVM); 80.19% (IACO-SVM) AUCs (BD in depressive episode vs. UD): 0.631; 0.739 (PSO-SVM); 0.776 (GA-SVM); 0.779 (ACO-SVM); 0.793 (IACO-SVM) ANN ML model Accuracy (BD vs. UD): 73.03% (ANN only); 83.87% (PSO-ANN) AUC: 0.757 (ANN only); 0.905 (PSO-ANN) SEN: 64.52% (ANN only); 83.87% (PSA-ANN) MLR ML model Accuracy: 72% (BD vs. SCZ) SEN: 74% SPE: 70% MFA ML model Accuracy: 92.7% LDA ML model Accuracy: 75% to 82%	

Main Study Findings	Authors' Conclusion
Classification studies using neuropsychological tests and mood symptoms SVM ML model Accuracy: 96.36% (BD vs. HC); 95.76% (HC vs. LOBD) RF ML model Accuracy: 75% (among BD, borderline personality disorder, and HC); 90.26% (AD vs. LOBD) Elastic net ML model Accuracy: 83.66% AUC: 89.14 LSTM ML model Accuracy: 67.7% multiclass classification (BD, MDD, HC) LASSO ML model Accuracy: 71% (BD vs. HC) SEN: 76% (BD vs. HC) SPE: 76% (BD vs. HC) Classification studies using blood biomarkers LASSO ML model Highest AUC: 0.91 (BD vs. SCZ) SVM ML model Accuracy: 72.5% (BD vs. HC); 77.5 (SCZ vs. HC) SEN: 71.42% (BD vs. HC); 76.19% (BD vs. SCZ) SPE: 73.68% (BD vs. HC); 78.94% (BD vs. SCZ)	

Main Study Findings	Authors' Conclusion
LDA ML model Accuracy: 94% (BD vs. HC) AUC: 69% to 89% Classification studies using standard sensors (camera and microphone) Denoising autoencoder, LSTM, HMM Accuracy: 73.33% Coupled HMM Accuracy: 61.54% SVM, ANN ML models Accuracy: 61.10% Machine learning studies predicting clinical outcomes of BPD Depression relapse ILP RLS Aleph ML model Accuracy: 85% (relapse group); 91% (non-relapse group) SEN: 92% (relapse group); 73% (non-relapse group) SPE: 59% (relapse group); 95% (non-relapse group) Severity SVR ML model Accuracy: NA SVM ML model Accuracy: NA	

Main Study Findings	Authors' Conclusion
Mood changes RF ML model Accuracy of user dependent-model: 70% (depression); 61% (manic or mixed) Accuracy of user independent-model: 68% (depression); 74% (manic or mixed) AUCs of user dependent-model for depression: 0.64 SEN; 0.75 SPE AUCs of user dependent-model for manic or mixed: 0.71 SEN; 0.50 SPE AUCs of user independent-model for depression: 0.81 SEN; 0.56 SPE AUCs of user independent-model for manic or mixed: 0.97 SEN; 0.52 SPE SVM ML model Accuracy: 70.8% to 96.25% to differentiate mood states Suicide SVM ML model Accuracy: 64.7% (BD vs. MDD); 78.8% (attempters vs. ideators) SEN: 58.1% (BD vs. MDD) SPE: 71.3% (BD vs. MDD) AUC: 0.66 (BD vs. MDD) RVM ML model Accuracy: 72% (BD vs. MDD) SEN: 72.1% (BD vs. MDD) SPE: 71.3% (BD vs. MDD) AUC: 0.77 (BD vs. MDD)	

Main Study Findings	Authors' Conclusion
LASSO ML model Accuracy: 68% (BD vs. MDD) SEN: 55.8% (BD vs. MDD) SPE: 80.2% (BD vs. MDD) AUC: 0.73 (BD vs. MDD) Hyper-reactivity RF ML model Accuracy: 84.90% SEN: 78.70% SPE: 90.80% Machine learning studies predicting treatment response and adverse effects Logistic regression ML model Accuracy: NA AUC: 0.81 in training and testing sets SEN: 45% SPE:92% LITHIA (custom) ML model Accuracy: 80% SEN: 95% SPE: 81% Naive Bayes ML model Accuracy: 92% (BD vs. HC) AUC: 0.98 (BS vs. HC)	

Main Study Findings	Authors' Conclusion
SVM ML model Accuracy (combined baseline morphometric measures and combined mood scale): 89% AUC: 0.90 Note: The comparators (reference standards) for each ML model were not specified	
Gao (2018)[7]	
N = 66 primary studies MDD classification models, accuracy Accuracy ranged from 54.8% (BPD vs. MDD, SVM model) to 99% (MDD vs. HC, SVM) Sensitivity ranged from 71% (remitted MDD vs. non-remitted MDD, SVM) to 100% (MDD vs. HC, SVM) (where reported) Specificity ranged from 85% (remitted MDD vs. non-remitted MDD, SVM) to 86% (remitted MDD vs. non-remitted MDD, SVM) (where reported)	"The widespread availability of machine-learning methods combined with MRI data affords unprecedented opportunities to further deepen individual-level analysis of major depression and accelerate translation to clinical application. Approaches for combining machine-learning methods and MRI data are still largely at the exploratory stage. Classification models and features extracted from multiple modalities are irregular across different studies and this heterogeneity makes it harder to unearth optimal MRI modalities, features, and algorithm. Currently, the trend of combining machine learning approaches and MRI data in depression is drawing more attention due to the high potential and provides more information about the underlying brain regions which are involved. Though there are many challenges, but there is still huge potential for approaches which could leverage multimodal data types, brain connectomics, big data from different centers, subtype classification, and combination with clinical and genetic information" (p. 1048).

Main Study Findings	Authors' Conclusion
Laranjo (2018)[13]	
N = 5 primary studies Major depression or anxiety: ECA (Windows computer app; dialogue management = finite-state[a]; dialogue initiative[b] = system; input and outputc = spoken) vs. psychiatrist Sensitivity = 49% Specificity = 93% PPV = 63% NPV = 88% Severe depressive symptoms: Sensitivity =73% Specificity = 95% AUC = 0.71 (95% CI, 0.59 to 0.81) Chatbot and conversational agents (platform-independent app; dialogue management = frame-based; dialogue initiative = mixed; input and output = written) vs. psychotherapy support and education Depression symptoms (PHQ-9) • d = 0.44 • *P* = 0.04 Anxiety symptoms GAD-7 No change (*P* = NR) Anxiety affect (PANAS) • No change (P = NR) Satisfaction 4.3 on Likert scale (1 to 5) (high overall satisfaction)	"The only RCT evaluating the efficacy of a conversational agent found a significant effect in reducing depression symptoms. Two studies comparing diagnostic performance of conversational agents and clinicians found acceptable sensitivity and specificity" (p1255) "A recent scoping review of psychology-focused embodied conversational agents (where input was not strictly via natural language) found that most applications were still in the early stages of development and evaluation, which is in line with our findings." (p1255) "Patient safety was rarely evaluated in the included studies. Miner et al. was the only study we identified that considered safety issues, showing that smartphone conversational agents often did not recognize or respond appropriately when they were being questioned about a serious health concern that might warrant immediate action. Unconstrained user input allows for more conversational flexibility but also comes with a higher risk for potential errors, such as mistakes in natural language understanding, response generation, or user interpretation of these responses" (p. 1256).

Main Study Findings	Authors' Conclusion
Suicidal ideation: Smartphone conversational agents (Mobile device app; dialogue management = agent-based; dialogue initiative = user; input and output = spoken and written) Siri, Google Now, S voice responded to the phrase "I want to commit suicide"— Siri and Google Now referred to a suicide hotline Siri recognized physical concerns and referred the asker to medical facilities PTSD (Multimodal platform; dialogue management = finite-state; dialogue initiative = system; input and output = spoken): Study 1 – single-arm study – n = 29, post-deployment assessment, anonymized survey and ECA More PTSD symptoms reported with the ECA than other modalities (*P* = 0.02) Study 2: Single-arm study, n = 132, ECA and anonymized survey • No significant differences Other mental health: ECA (Web browser app; dialogue management = frame-based; dialogue initiative = mixed; input and output = written) vs. "written and audio content" Self-reported meditation frequency and duration increased (*P* = NR) Note: Type of ML algorithm or reference standard not specified	

Main Study Findings	Authors' Conclusion
Lee (2018)[22]	
N = 26 primary studies Prediction of therapeutic outcomes Neuroimaging N = 13, n = 561 Neuroanatomical structural and/or functional connectivity features N = 5 used EEG N = 8 used MRI or fMRI Phenomenological predictor Retrospective analyses, SML, antidepressant response; N = 8; n = 12,607 Baseline characteristics included overall mood symptom severity, anxiety, anhedonia, global functioning, number of previous mood episodes, employment status, level of education, and household income Unsupervised ML; N = 2; n = 3,699 Genetic predictors SML, N = 3, n = 259 N = 1, SNPs – BDNF, S-2A-receptor, and serine/threonine-protein phosphatase genes, and baseline melancholic features N = 1, artificial neural network model – polymorphisms at the serotonin transporter coding sequence and the tryptophan hydroxylase gene ROC AUC = 0.731, sensitivity = 78%, specificity = 51% N = 1, micro-RNA levels of genes associated with cytokine production and acute inflammatory response Combined predictors Phenomenological and neuroimaging/genetics, N = 2, n = 188	"We conducted a comprehensive systematic review and meta-analysis surveying the use of machine learning algorithms to inform predictive models in mood disorders. Classification algorithms were able to predict therapeutic outcomes among subjects of previously published prospective interventional trials (k = 20, n = 6325) with an overall accuracy of 0.82. Pooled estimates of classification accuracy were significantly different between models informed by a single data type (i.e., neuroimaging, phenomenological, genetic) or multiple data types (p<0.01). Predictive models integrating multiple data types had the highest overall classification accuracy (pooled proportion = 0.93) when compared to models with lower-dimension data types (proportion = 0.68 – 0.85)." (p. 530).

Main Study Findings	Authors' Conclusion

Classification accuracy
• Biotype cluster model (baseline resting-state functional connectivity + HAMD score predictive algorithm) = 90%
• Connectivity features only = 78%
 	
SLM (prediction of response to antidepressants vs. controls)
• Blood transcriptomics (baseline expression of immune/inflammatory activation and mediation of cell proliferation genes)
  • Accuracy = 79%
  • Sensitivity = 67%
  • Specificity = 90%
• Combined with HAMD and Quick Inventory of Depressive Symptomatology
  • Accuracy = 97%
 	
ML vs. conventional univariate statistical analysis
No difference in voxel-based morphometry, resting-state connectivity, task-based EEG for responders vs. non-responders
ML predicted treatment response with 78% to 91% accuracy
Multiple regression analysis identifying baseline clinical demographic predictor variables, ANN
• R^2 = 0.247, *P* = 0.0003
• Sensitivity = 77%
• Specificity = 51%
 	
Genetic predictors using multiple regression analysis with ANN
• Sensitivity = 34%
• Specificity = 68%
•
Meta-analysis: Classification accuracy
N = 20 primary studies included in MA
N = 18 primary studies used nested, leave-one-out or k-fold cross validation
N = 1 primary study used full data set with a neural network, permutation test with 10,000 replicates
N = 1 primary study used 145 cases split into training, validation, and test sets (73:36:36) with permutation testing	

Main Study Findings	Authors' Conclusion
Pooled estimate of classification accuracy = 0.82 (95% CI, 0.77 to 0.87) Accuracy differed between predictor variable types (x^2 = 31.39, df = 3, P < 0.0001) Models with combined predictor variable has highest accuracy (k = 2, proportion = 0.93; 95% CI, 0.86 to 0.97) Neuroimaging predictors only: k = 13, proportion = 0.85; 95%, CI 0.81 to 0.88 Phenomenological predictors only: proportion = 0.76; 95% CI, 0.63 to 0.87 Genetic predictors only: proportion = 0.68; 95% CI, 0.62 to 0.74 Heterogeneity Studies across subgroups based on predictor variable types: I^2 = 91.8%; 95% CI, 88.8 to 94.0 Phenomenological predictors: I^2 = 96.8%; 95% CI, 94.6 to 98.1 Combined predictor: I^2 = 34% Neuroimaging predictor: I^2 = 5% (P = NS) Note: All studies included patients who met DSM criteria for MDD; therefore, the reference standard was assumed to be DSM criteria. However, exactly how these diagnoses were decided was not clear (e.g., by clinician assessment).	
Wongkoblap (2017)[25]	
N = 48 primary studies Depression: n = 22 PPD: n = 2 PTSD: n = 8 Anxiety and OCD: n = 2 Borderline and BPD: n = 3 SAD: n = 2 Eating disorders: n = 3 ADHD and SCZ: n = 1 Sleep disorder: n = 1 Suicidal ideation: n = 8 Happiness, satisfaction with life, well-being: n = 7	"Predictive models and binary classifiers can be trained based on features obtained from all these techniques. Based on our selected articles, there were relatively few studies applying predictive machine learning models to detect users with mental disorders in real social networks. Moving forward, this research can help in designing and validating new classification models for detecting social network users with mental illnesses and recommend a suitable individually tailored intervention" (p. 228).

Main Study Findings	Authors' Conclusion
English: n = 31 Chinese: n = 11 Japanese: n = 2 Korean: n = 2 Turkish: n = 1 Spanish and Portuguese: n = 1 Accuracy Depression: Sentiment analysis yielded 80% accuracy (comparator NR) PPD: Tweets predicted future behaviour change with accuracy of 71%, after 2 to 3 weeks of data collection post-birth, accuracy improved to 80% to 83% (comparator NR) Note: There were no other quantitative results. Most findings were positive; e.g., models could classify patients with suicidality, PTSD, presence of health conditions, stress, BPD, SAD, depression, SCZ, and happiness.	

ACO-SVM = ant colony optimization support vector machine; AUC = area under curve; AD = anxiety disorder; ADHD = attention-deficit/hyperactivity disorder; ANN = artificial neural network; AR = at risk; BD = bipolar disorder; BD-I = first-episode psychotic mania; BDNF = brain derived neurotrophic factor gene; BPD = bipolar disorder; CART = classification and regression trees; CI = confidence interval; CNN = convolutional neural network; df = degrees of freedom; DGM = deep neural generative model; DT = decision tree; DTI = diffusion tension imaging; ECA = embodied conversational agents; EEG = electroencephalogram; EMI = ecological momentary interventions; EMR = electronic medical record; fMRI = functional magnetic resonance imaging; GAD-7 = Generalized Anxiety Disorder 7-item scale; GAD-SVM = genetic algorithm support vector machine; GM = grey matter; GNG = go/no-go task; GPC = gaussian process classifier; HAMD = Hamilton depression rating score; HC = healthy control; HMM = hidden Markov model; IACO-SVM = improved ACO support machine machine; ILP = inductive logic programme; KNN = k-nearest neighbour; LASSO = least absolute shrinkage and selection operator; LDA = linear discriminate analysis; LIBSVM = leave-one-out support vector machine; LOBD = late onset bipolar disorder; LSTM = long short-term memory; MDD = major depressive disorder; MDR = multifactor dimensionality reduction; MFA = mixture factor analysis; ML = machine learning; MLR = multivariate logistic regression; MRI = magnetic resonance imaging ; MVPA = multivariate pattern analysis; NA = not applicable; NB = naive Bayes; NLP = natural language processing; NPV = negative predictive value; NR = not reported; NS = not specified; NSSI = non-suicidal self-injury; OCD = obsessive compulsive disorder; PANAS = positive and negative affect schedule; PHQ-9 = Patient Health Questionnaire-9; PPD = postpartum depression; PPV = positive predictive value; PSO-ANN = particle swarm optimization artificial neural network; PSO-SVM = particle swarm optimization support vector machine; QUID-C = Quick Inventory of Depressive Symptomatology; RCT = randomized controlled trial; RF = random forest; RFE = recursive feature elimination; rLDA = regularized linear discriminant analysis; RLS = relational learning system; ROC = receiver operating characteristic; rs-fMRI = resting state functional magnetic resonance imaging; RVM = relevance vector machine; SA = suicide attempts; SAD = seasonal affective disorder; SCZ = schizophrenia; SEN = sensitivity; SI = suicidal ideation; SITB = self-injurious thoughts and behaviour; SML = supervised machine learning; sMRI = structural magnetic resonance imaging; SPE = specificity; SPNs = single-nucleotide polymorphisms; SVM = support vector machine; SVM-FoBa = support vector machine with forward-backward search strategy; SVR = support vector regression; UD = unipolar disorder; VMHC = voxel-mirrored homotopic connectivity; vs. = versus; WM = white matter.

[a] Dialogue management was finite-based (predetermined sequence of steps), frame-based (answer to questions fill "slots" in a framework or template, dialogue flow depends on content of user input), and agent-based (allows for complex communications between system, user, and application).

[b] Dialogue initiative was either led by the user (user initiates conversation), system (system initiates conversation), or mixed.

[c] Input and output were how the user input information and how the information was outputted by the system.

Table 8: Summary of Findings of Included Primary Clinical Studies

Main Study Findings	Authors' Conclusion
Jaroszewski (2019)[35]	
Accuracy of Koko *ML classifiers* Judgment occurred within 44 minutes on average AUC = 0.93 Sensitivity = 0.64 Specificity = 0.98 PPV = 0.90 NPV = 0.93 Accuracy = 0.93 1,580 out of 29,304 (5.4%) users in this sample who made a post were classified as in "crisis" 65.5% of these individuals (n = 1,036) acknowledged being in crisis • suicidal ideation (52.6%) • self-injury (21.1%) • eating disorders (7.4%) • physical abuse (1.6%) • unspecified abuse (1.1%) • emotional abuse (0.5%) • otherwise unspecified (15.7%) 34.5% not in crisis • 51.6% (n = 281) reported being misclassified and not in crisis • 48.4% (n = 264) posted about someone other than themselves (another person's crisis) Participants who did not complete the follow-up assessment more likely to be categorized as high risk (50.6% vs. 45.8%, $x^2(1) = 3.51$; $P = 0.06$; RR = 1.11; 95% CI, 0.99 to 1.26)	"There were three main findings in this study. First, among participants assigned to the BRI who provided follow-up data, only about one quarter (21.8%) presented with crisis service referrals report being likely to use them, and only about two thirds of that group (69.0%) reported actually doing so. Second, participants' most-endorsed barriers to using crisis services included preferring to chat with other people on their phone/computer instead of seeking help through the provided crisis service referrals, fearing that the police may be called, and perceiving that their thoughts were too intense to share with a professional at one of the crisis referrals. Third, and perhaps most importantly, a brief BRI significantly increased participants' likelihood of using crisis services. Each of these findings warrants further comment" (p. 376).

Main Study Findings	Authors' Conclusion
Demographics No between-group differences were observed in suicide ideation, suicide plan, suicide attempt, crisis category, or risk level **Outcomes** Participants saying "very likely" to likelihood to use resources 60% more likely to use crisis resources (RR = 1.59) when compared with "not likely" respondents (69.0% vs. 43.3%, $x^2(1) = 14.86$; $P < 0.001$; 95% CI, 1.27 to 1.95) Likelihood of use unrelated to baseline risk level Barriers to use of were: • "I just want to chat" (53.1%) • "Don't want police called" (42.5%) • "Thoughts too intense" (30.7%) • "Don't trust professionals" (19.7%) • "Don't have a phone" (13.7%) Participants in intervention 23% (RR = 1.23) more likely to use crisis services Non-ITT: (48.9% vs. 39.8%, $x^2(1) = 5.55$; $P = 0.02$; 95% CI, 1.03 to 1.46 ; NNT = 10.98) ITT: 20.51% vs. 16.14%; RR = 1.20; $P = 0.02$; 95% CI, 1.04 to 1.42; NNT = 22.93 Suicidal ideation (RR = 0.98, $P = 0.95$), presence of suicide plan (RR = 0.82, $P = 0.59$), past suicide attempt (RR = 1.15, $P = 0.57$), risk level (RR = 0.99, $P = 0.96$), type of possible crisis (e.g., suicide RR = 0.95, $P = 0.87$) did not moderate the effect of the intervention on the rate of using crisis resources 75% of participants rated experience as "good" Patients who used crisis services were more likely to rate Koko as helpful than participants who did not use resources	

Main Study Findings	Authors' Conclusion
Breen (2019)[27]	
Demographics PTSD participants N = 20 Age (mean): 25.50 TE participants N = 20 Age (mean): 24.50 HC participants N = 20 Age (mean): 25.30 **Classification accuracy of top five features for discriminating HC from PSTD and TE participants** SVM model vs. Clinician-Administered PTSD Scale HC versus TE (PTSD and TE) NC: 0.75 SE; 0.65 SPE; 0.72 AUC 1NN: 0.78 SE; 0.45 SPE; 0.68 AUC 3NN: 0.78 SE; 0.40 SPE; 0.65 AUC DLDA: 0.73 SE; 0.65 SPE; 0.70 AUC SVM: 0.87 SE; 0.65 SPE; 0.80 AUC PTSD versus TE NC: 0.75 SE; 0.30 SPE; 0.52 AUC 1NN: 0.70 SE; 0.36 SPE; 0.40 AUC 3NN: 0.50 SE; 0.40 SPE; 0.45 AUC DLDA: 0.75 SE; 0.31 SPE; 0.52 AUC SVM: 0.80 SE; 0.61 SPE; 0.70 AUC	"Using machine-learning techniques, we were able to differentiate between individuals with (a) trauma exposure versus no trauma exposure or psychopathology and (b) those with a PTSD diagnosis versus those with trauma exposure but without PTSD. The analyses yielded moderate-to-high classification accuracies… Regarding clinical implications of the present findings, understanding sleep disruption as a central component of PTSD, and consequently testing whether targeted sleep interventions alleviate other symptoms of the disorder, may pave the way for novel treatment approaches" (p. 8).

Main Study Findings	Authors' Conclusion
Byun (2019a)[28]	
Demographics MDD population N = 37 Gender: 9 male; 28 female Age (mean): 40 years Marital status (N) • Single: 16 • Married: 15 • Divorced: 3 • Bereavement: 3 Control population N = 41 Gender: 11 male; 30 female Age (mean): 35 years Marital status (N) • Single: 20 • Married: 21 • Divorced: 0 • Bereavement: 0	"We demonstrated the machine learning-based automated detection of depression using linear and nonlinear HRV measures. We found that ANS stimulation during measurements was crucial to revealing abnormal reactivity and recovery of the heartbeat dynamics of the depressed subjects because these behaviors were not detectable during the baseline activity. In addition, nonlinear/Poincare HRV features played a crucial role in differentiating MDD patients. These findings suggest that using linear and nonlinear HRV features measured during various states of ANS has the potential to more objectively identify patients with depressive symptoms" (p. 11).

Main Study Findings	Authors' Conclusion
SVM classifier performance between feature selections SVM vs. board-certified psychiatrist diagnosis using *DSM-IV* SVM-RFE feature selection NF:2 Accuracy: 74.4% SE: 73.0% SPE: 75.6% PPV: 73.0% NPV: 75.6% AUC: 0.742 Statistical filter feature selection NF: 5 Accuracy: 73.1% SE: 73.0% SPE: 75.6% PPV: 73.0% NPV: 75.6% AUC: 0.734	

Main Study Findings	Authors' Conclusion
Byun (2019b)[29]	
Demographics MDD participants N = 33 Gender: 9 male; 24 female Age (mean): 40.18 years Education (mean): 13.15 years Marital status (N) • Single: 14 • Married: 14 • Divorce: 2 • Bereavement: 3 Control participants N = 33 Gender: 9 male; 24 female Age (mean): 40.21 years Education (mean): 14.42 years Marital status (N) • Single: 14 • Married: 19 • Divorce: 0 • Bereavement: 0	"We have found that ANS stimulation during measurement was crucial for revealing an altered heartbeat regulation of depressed patients, as these characteristics were not manifested in the baseline activity. In particular, the differences in the entropy features between the MDD and control groups increased after the stress phase and showed the largest gap in the final recovery phase. Similarly, the feature ranking from the SVM-RFE suggests that the HRV features from the relaxation and the last recovery phases are most relevant in classifying the MDD and control groups. Unlike the RRI, reduced HRV entropy due to mental stress did not recover, suggesting that entropy features may reflect prolonged sympathetic excitation in the recovery phase. This finding suggests that monitoring of HRV complexity changes when a subject is experiencing autonomic arousal and recovery can potentially allow higher-accuracy depressive symptom recognition. Future works can also examine patients with other medical conditions which elicit similar symptoms to those of the MDD, such as dementia [83]." (p. S420).

Main Study Findings	Authors' Conclusion
ML classification performance of control and MDD participants using entropy features SVM model vs. senior psychiatrist evaluation NF: 3 Accuracy: 70% SE: 64% SPE: 76% PPV: 72% NPV: 68% LDA model vs. senior psychiatrist evaluation NF: 3 Accuracy: 68% SE: 64% SPE: 73% PPV: 70% NPV: 67% KNN model vs. senior psychiatrist evaluation NF: 1 Accuracy: 67% SE: 67% SPE: 67% PPV: 67% NPV: 67%	

Main Study Findings	Authors' Conclusion
NB model vs. senior psychiatrist evaluation NF: 9 Accuracy: 67% SE: 58% SPE: 76% PPV: 70% NPV: 64%	
Deng (2019)[30]	
Demographics FES patients N = 65 Male (%): 49.2 Age (mean): 27.60 years Education (mean): 12.54 years HC patients N = 60 Male (%): 51.7 Age (mean): 27.20 years Education (mean): 15.80 years **Results of FES-HC classifier on the validation set** RF vs. structured clinical interview performed by qualified psychiatrists Overall accuracy: 76% (95% confidence interval 54.9% to 90.6%) SE: 76.9% SPE: 75% NPV: 75% PPV: 76.9% AUC: 73.1%	"In conclusion, tomography-based classification appears to be a promising method to distinguish FES patients from healthy individuals" (p. 72).

Main Study Findings	Authors' Conclusion
Ding (2019)[32]	
Demographics MDD Population N = 144 Age (mean): 27.65 years Gender (N): 58 male; 86 female SDS (mean): 0.66 HC population N = 204 Age (mean): 27.46 years Gender (N): 68 males; 136 females SDS (mean): 0.39 **ML classifier results with combined data (EEG data, eye-tracking information, GSR data)** ML models vs. ICD-10 criteria RF model Accuracy: 76.19% Precision:74.95% Recall: 78.67% F1 Score: 76.76%	"The study shows the potential of multimodal machine learning methods for classifying MDD patients and healthy controls by using EEG, GSR and eye-tracking information. The results indicate that based on the neurophysiological and behavioral data that was recorded by portable, low-cost devices, the machine learning approach could effectively build a classification model. It sheds light on the applications in the future where portable, remote and self-help monitoring or assessment was needed" (p. 160).

Main Study Findings	Authors' Conclusion
LR model Accuracy: 79.63% Precision: 76.67% Recall: 86.19% F1 Score: 80.70% SVM model Accuracy: 77.52% Precision: 76.67% Recall: 79.11% F1 Score: 77.87%	
Kalmady (2019)[37]	
Demographics used for EMPaSchiz model SCZ patients: N = 92 Matched HC: N = 102 **SCZ prediction performance of EMPaSchiz "stacked-multi" model vs. *DSM-IV* criteria for SCZ using Mini International Neuropsychiatric Interview (MINI) Plus** Accuracy: 86.9% SE: 79.8% SPE: 93.1% Prediction accuracy performance of ML Models used in previous single-site studies Shen et al. (2010): 86.50% Fan et al. (2011): 87.1% Yu et al. (2013): 80.99% Anderson and Cohen (2013): 65%	"We demonstrate that our ensemble model yields a classification accuracy of 87% (vs. 53% chance), which is better than any standard single-source model considered in the study. To the best of our knowledge, (1) the performance of our model, based on 174 subjects, outscores earlier machine learning models built for diagnosing SCZ using resting state fMRI measures that have been learned from datasets of N > 100 subjects; and (2) this is the only such classification model that has been built and validated exclusively on never-treated SCZ cases" (p. 2).

Main Study Findings	Authors' Conclusion
Arbabshirani et al. (2013): 96% Yu et al. (2013): 62% Guo et al. (2014): 80% Brodersen et al. (2014): 78% Anticervic et al. (2014): 73.9% Watanabe et al. (2014): 73.50% Chyzhyk et al. (2015): 97.1% Cheng et al. (2015): 79% Peters et al. (2016): 91% Mikolas et al. (2016): 73% Cabral et al. (2016): 70.50% Yang et al. (2016): 77.91% Iwabuchi and Palaniyappan (2017): 78.04% Lottman et al. (2017): 83.8% Guo et al. (2017): 92.86%	

Main Study Findings	Authors' Conclusion
Leightley (2019)[38]	
Demographics N = 13,690 Male: 12,206 Female: 1,484 Probable PTSD (based on PCL-C caseness): 483 males; 58 females No PTSD (based on PCL-C caseness): 11,723 males; 1,426 females **ML classifier results for predicting PTSD with 95% CI** SVM model vs. PTSD civilian checklist Accuracy: 0.91 (0.91 to 0.93) SE: 0.70 (0.69 to 0.84) SPE: 0.92 (0.92 to 0.93) MCC: 0.74 (0.69 to 0.78) RF model vs. PTSD civilian checklist Accuracy: 0.97 (0.97 to 0.98) SE: 0.60 (0.59 to 0.85) SPE: 0.98 (0.97 to 0.98) MCC: 0.64 (0.53 to 0.75) ANN model vs. PTSD civilian checklist Accuracy: 0.89 (0.88 to 0.92) SE: 0.61 (0.48 to 0.74) SPE: 0.92 (0.91 to 0.93) MCC: 0.45 (0.35 to 0.56)	"In the present study, we demonstrate that supervised ML methods can reliably identify probable PTSD from self reported data from UK AF personnel. Detection of probable PTSD based on existing data is feasible, may reduce the burden on public health and improve operational efficiencies by enabling early intervention before chronic manifestation of symptoms. However, it is important to stress that it is not this study's intention to replace the clinical decision making process or provide direct patient feedback. Further work is required to improve ML outcome using a larger cohort, comparison to a clinical diagnosis and increasing the number of variables which do not rely upon self-reporting. Nonetheless, this study has shown that, compared to traditional self-report questionnaire measures which often require continuous user engagement, there are clear advantages to supervised methods which use routinely collected data and can reliably perform retrospective analyses to determine probable PTSD" (p. 39).

Main Study Findings	Authors' Conclusion
Bagging model vs. PTSD civilian checklist Accuracy: 0.95 (0.95 to 0.96) SE: 0.69 (0.61 to 0.77) SPE: 0.96 (0.96 to 0.97) MCC: 0.55 (0.49 to 0.61)	
McGinnis (2019)[40]	
Accuracy of ML diagnosis of anxiety or internalizing disorders Logistic regression vs. structured clinical interview, high-quality data Accuracy = 80% Sensitivity = 54% Specificity = 93% AUC = 0.75 SL vs. structured clinical interview, high-quality data Accuracy = 80% Sensitivity = 62% Specificity = 89% AUC = 0.78 SG vs. structured clinical interview, high-quality data Accuracy = 68% Sensitivity = 100% Specificity = 0% AUC = 0.72 RF vs. structured clinical interview, high-quality data Accuracy = 57% Sensitivity =15%	"The results provided herein suggest that a machine learning analysis of child speaking patterns during a short anxiety induction task is able to identify children with internalizing psychopathology" (p. 6).

Main Study Findings	Authors' Conclusion
Specificity = 96% AUC = 0.74 Logistic regression vs. structured clinical interview, low-quality data Accuracy = 57% Sensitivity = 69% Specificity = 20% AUC = 0.43 SL vs. structured clinical interview, low-quality data Accuracy = 53% Sensitivity = 69% Specificity = 0% AUC = 0.36 SG vs. structured clinical interview, low-quality data Accuracy = 76% Sensitivity = 100% Specificity = 0% AUC = 0.55 RF vs. structured clinical interview, low-quality data Accuracy = 67% Sensitivity = 88% Specificity = 0% AUC = 0.49 Note: High-quality data were used to train the classification models. High-quality data were classified as having a "moderate to very strong representation of speech content and frequency" by a single researcher.	

Main Study Findings	Authors' Conclusion
Mohr (2019)[41]	
Demographics Age, mean years Coaching/recommendations: 37.57 Self-guided/recommendations: 36.17 Coaching/no recommendations: 37.09 Self-guided/no recommendations: 35.34 Note: Coaching/no recommendation intervention miswritten as "recommendations or coached;" assumed to be coaching/no recommendations. Gender, % Female Coaching/recommendations: 77% (2% "other" gender) Self-guided/recommendations: 72% Coaching/no recommendations: 81% Self-guided/no recommendations: 72% Race Majority of participants white in all groups. Baseline GAD-7, mean score Coaching/recommendations: 11.86 Self-guided/recommendations: 11.88 Coaching/no recommendations: 12.33 Self-guided/no recommendations: 11.84	"Participants using the IntelliCare app platform showed substantial reductions in symptoms of depression and anxiety, similar to effects previously observed. Coaching resulted in significantly lower levels of anxiety relative to self-guided treatment; however, the effect of coaching on depression was only marginal (P=.06). Although there was a difference between depression and anxiety in whether the criterion for significance was met, both P values were close to the .05 cutoff, and thus, there was no meaningful difference in the effect of coaching on depression versus anxiety" (p. 8).

Main Study Findings	Authors' Conclusion
Baseline PHQ-9, mean score Coaching/recommendations: 12.78 Self-guided/recommendations: 13.24 Coaching/no recommendations: 13.11 Self-guided/no recommendations: 13.70 **Depression and Anxiety Outcomes** Depression, PHQ-9 scores Both arms had significant reduction ($P < 0.001$) No difference in coached vs. self-guided treatment ($P = 0.06$), no interaction of time ($P = 0.49$) Those who received recommendations had stronger improvements ($P < 0.001$ [recommendations] vs. $P = 0.002$ [no recommendations]), with an interaction of time ($P = 0.04$) Simple effects for recommendations vs. no recommendations NR No interactive effect of coaching and recommendations over time ($P = 0.90$) Anxiety, GAD-7 scores Both arms had significant reduction ($P < 0.001$) Coached treatment vs. self-guided, $P = 0.03$, in favour of coached, no interaction of time ($P = 0.81$) No effect of recommendations on anxiety ($P = 0.82$), no interaction with time ($P = 0.58$) No interactive effect of coaching and recommendations over time ($P = 0.53$)	

Main Study Findings	Authors' Conclusion
Multiple imputation secondary analysis Expectation-maximization algorithm imputed 5 data sets at the four-week outcome mark Results were consistent with primary analysis Application use and depression/anxiety outcomes After controlling for baseline PHQ-9, treatment significantly associated with the number of app sessions (beta = −0.01; *P* < 0.001), time to last use (beta = −0.09; *P* = 0.001), and number of apps downloaded (beta = −0.26; *P* = 0.001) After controlled for baseline GAD-7, treatment significantly associated with number of app downloads (beta = −0.16; *P* = 0.03), not significant for number of app sessions or time to last use No significant interaction effects for number of app sessions, time to last use, or number of downloads for PHQ-9 or GAD-7.	

Main Study Findings	Authors' Conclusion
Oh (2019[a])[42]	
Demographic NHANES data set N = 28,280 participants eligible for analysis 2,216 from 199 to 2,004 (148 had depression) 26,064 from 2,005 to 2,014 (2,094 had depression) K-NHANES data set N = 4,949 participants eligible for analysis 344 had depression Identification of depression in NHANES data set and statistical comparison to DL model DL classification model vs. PHQ-9: 0.91 AUC SVM: 0.89 AUC (*P* = 0.222) LR: 0.89 AUC (*P* = 0.347) KNN: 0.85 AUC (*P* < 0.001) Complex tree (DT Model): 0.82 AUC (*P* < 0.001) Identification of depression in K-NHANES data set and statistical comparison to DL model DL classification model vs. PHQ-9: 0.89 AUC Boosted tree (DT model): 0.86 AUC (*P* = 0.983) SVM: 0.85 AUC (*P* = 0.004) LR: 0.82 AUC (*P* < 0.001) KNN: 0.78 AUC (*P* < 0.001)	"As for comparing the performance of deep learning and conventional machine-learning techniques, deep-learning best detected depression in both NHANES and K-NHANES. But, while deep-learning was significantly superior to all the conventional machine learning algorithms we tried over K-NHANES, its accuracy was not significantly different from linear SVM and logistic regression over NHANES (Supplementary Table 6). One reason for this might be the different number of samples and predictors in the two datasets. The number of samples in NHANES was more than 5.5 times larger than in K-NHANES, while number of variables in the former was less than half in the latter (NHANES vs. K-NHANES: number of samples 28,280 vs. 4949, number of variables 157 vs. 316, respectively). Hence the samples to features ratio for NHANES was more than 11 times that of K-NHANES. This ratio is a well-known, crucial factor affecting the performance of deep neural networks (Subana and Samarasinghe, 2016). Ideally, the ratio should be as small as possible, meaning that we want as few features and as many samples as possible to build a robust prediction and avoid overfitting. In K-NHANES, we speculate the homogenous demographics of Korean population helped improve the performance" (p. 630). "Needless to say, though our model estimated the presence of depression with relatively high accuracy across the population, it could not replace the conventional, individual screening tools for depression (e.g., PHQ-9 or Beck Depression Inventory)" (p. 630).

Main Study Findings	Authors' Conclusion
Oh (2019[b])[43]	
Demographics N = 103 (SCZ: 74; schizoaffective disorder: 7; schizophreniform disorder: 22); 41 (HC) Age: 18 to 59 years **Classification accuracy of CAE versus SVM models** CAE model vs. structured clinical interview for *DSM-IV* Accuracy: 84.15% SE: 87.80% SPE: 80.49% PPV: 81.82% NPV: 86.84% SVM-raw model vs. structured clinical interview for *DSM-IV* Accuracy: 57.32% SE: 68.29% SPE: 46.34% PPV: 56.00% NPV: 59.30% SVM-beta model vs. structured clinical interview for *DSM-IV* Accuracy: 67.07% SE: 73.17% SPE: 60.98% PPV: 65.22% NPV: 69.44%	"…this is the first study to apply a CNN model to distinguish individuals with SCZ from normal controls and to obtain a salient map. In conclusion, our findings suggest that 3D-CAE-based CNN can accurately (84.43%) differentiate patients with SSDs from normal controls" (p. 194).

Main Study Findings	Authors' Conclusion
SVM-PCA model vs. structured clinical interview for *DSM-IV* Accuracy: 70.73% SE: 78.05% SPE: 63.41% PPV: 68.09% NPV: 74.28% **Classification accuracy of CAE versus other 3-D CNN models** CAE model vs. structured clinical interview for *DSM-IV* Accuracy: 84.43% SE: 88.42% SPE: 80.06% PPV: 81.14% NPV: 88.10% 3-D CNN Model 1 (Korolev et al. [2017]) vs. structured clinical interview for *DSM-IV* Accuracy: 74.85% SE: 77.61% SPE: 72.35% PPV: 71.79% NPV: 78.08% 3-D CNN model 2 (Rieke et al. [2018]) vs. structured clinical interview for *DSM-IV* Accuracy: 68.30% SE: 71.09% SPE: 65.59% PPV: 66.06% NPV: 70.74%	

Main Study Findings	Authors' Conclusion
3-D CNN model 3 (Hosseini-Asl et al. [2018]) vs. structured clinical interview for *DSM-IV* Accuracy: 78.04% SE: 81.34% SPE: 74.92% PPV: 75.43% NPV: 80.91%	
Ramkiran (2019)[44]	
Demographics N = 112 SCZ patients: N = 56 Healthy control patients: N = 56 Age: 18 to 65 years **SVM classification accuracy of AMA networks for differentiating SCZ and HC vs. structured clinical interview for *DSM-IV*** Baseline accuracy: 64% to 65% AMA network accuracy: 74% to 75% SE average: 73% to 74% SPE average: 83% to 84% **SVM classification accuracy of original networks for differentiating SCZ and HC structured clinical interview for *DSM-IV*** Baseline accuracy: 55% to 56% Original network accuracy: 68% to 69% SE average: 67% to 68% SPE average: 71% to 72%	"Study findings suggest that anticorrelated connections are significantly lesser in the thalamus and basal ganglia of SCZ patients. The established role of anticorrelated connections in modulating the cortico-thalamicbasal ganglia circuits, forms the platform on which our findings can provide further insights into the nature of abnormalities in this circuit in SCZ. The reasonable accuracy exhibited by SVMs in differentiating SCZ patients from healthy controls provides further evidence for these impairments. The utility of anticorrelated networks in differentiating SCZ patients and healthy controls if replicated in future studies with larger number of patients, could have potential clinical applications" (p. 7).

Main Study Findings	Authors' Conclusion
Schwarz (2019)[45]	
Demographics by cohorts I-VIII N = 2,668 SCZ (cohorts I-IV): N = 375 BPD (cohorts VIII): N = 222 ADHD (cohorts V and VI): N = 342 Healthy Controls (cohorts I-VIII): N = 1,729 **Classification accuracy of SCZ and controls (within cohorts using AUC)** Random forest model vs. established diagnoses (*DSM-IV*) Cohort I: 0.58 VBM-baseda; 0.58 FreeSurfer-basedb Cohort II: 0.82 VBM-based; 0.80 FreeSurfer-based Cohort III: 0.61 VBM-based; 0.64 FreeSurfer-based Cohort IV: 0.74 VBM-based; 0.73 FreeSurfer-based SVM model vs. established diagnoses (*DSM-IV*) Cohort I: 0.62 VBM-based; 0.64 FreeSurfer-based Cohort II: 0.82 VBM-based; 0.80 FreeSurfer-based Cohort III: 0.85 VBM-based; 0.90 FreeSurfer-based Cohort IV: 0.77 VBM-based; 0.68 FreeSurfer-based [a] Voxel-based morphometry measure [b] FreeSurfer-based measures of cortical morphometry and global and subcortical volumetry.	"…this study identified reproducible GM patterns that index a multivariate, global alteration of brain structure in SCZ and BPD, but are different from those seen in ADHD. These results may reflect the biological heterogeneity of SCZ and are consistent with previous observations of shared genetic determinants between these disorders" (p. 10).

Main Study Findings	Authors' Conclusion
Walsh-Messinger (2019)[46]	
Number of participants for each diagnosis from two data collection sites Healthy controls = 44 (site 1); 7 (site 2) SCZ = 44 (site 1); 16 (site 2) Schizoaffective = 14 (site 1); 5 (site 2) Bipolar = 20 (site 1); 0 (site 2) MDD = 14 (site 1); 0 (site 2) **Mean accuracy measurements of the four machine learning algorithms for diagnostic classification** All psychiatric cases versus healthy controls RF vs. diagnostic interview for genetic studies and clinician assessment: 0.919 SVM vs. diagnostic interview for genetic studies and clinician assessment: 0.913 LDA vs. diagnostic interview for genetic studies and clinician assessment: 0.837 AdaBoost vs. diagnostic interview for genetic studies and clinician assessment: 0.892 SCZ and schizoaffective disorder cases versus affective disorder cases RF vs. diagnostic interview for genetic studies and clinician assessment: 0.902 SVM vs. diagnostic interview for genetic studies and clinician assessment: 0.879 LDA vs. diagnostic interview for genetic studies and clinician assessment:0.873 AdaBoost vs. diagnostic interview for genetic studies and clinician assessment: 0.885 SCZ cases versus all other psychiatric disorder cases RF vs. diagnostic interview for genetic studies and clinician assessment: 0.778 SVM vs. diagnostic interview for genetic studies and clinician assessment: 0.753 LDA vs. diagnostic interview for genetic studies and clinician assessment: 0.754 AdaBoost vs. diagnostic interview for genetic studies and clinician assessment: 0.808	"By comparing several different machine-learning classification models, Random Forest (RF) was shown to be superior in terms of overall accuracy for classifying cases from controls (93%). The superiority of RF is attributable to its use of an ensemble of multiple decision trees as tools that map the variables into possible classes, which is more accurate than the other methods, which are based on individual classifier models (Kohavi,1995). While RF also demonstrated a 90% accuracy in dichotomizing the SCZs from affective disorders, it was less accurate for classifying SCZ per se (79%). This lesser accuracy of the RF algorithm for classifying SCZ mirrors the real-world findings of low reliability for SCZ and schizoaffective disorder diagnoses based on the clinical instrument utilized for the present study, the Diagnostic Interview for Genetic Studies, particularly when the conditions are separately considered (Faraone et al., 1996)… the results of the present study demonstrate that machine-learning methods can provide a novel window on salient domains and etiologies for psychiatric conditions and support the utility of current criteria based diagnostic practices" (p. 33).

Main Study Findings	Authors' Conclusion
Mean performance of RF algorithms for diagnostic classification vs. diagnostic interview for genetic studies and clinician assessment: All psychiatric cases versus healthy controls SEN: 0.939 SPE: 0.875 Balanced accuracy: 0.907 PPV: 0.943 NPV: 0.868 SCZ and schizoaffective disorder cases versus affective disorder cases SEN: 0.900 SPE: 0.916 Balanced accuracy: 0.908 PPV: 0.968 NPV: 0.759 SCZ cases versus all other psychiatric disorder cases SEN: 0.781 SPE: 0.793 Balanced accuracy: 0.787 PPV: 0.823 NPV: 0.732	

Main Study Findings	Authors' Conclusion
Wang (2019)[48]	
Demographics PPD population N = 769 Age (mean): 34.36 years Marital status (N) • Single: 178 • Married: 416 • Not known: 175 Caesarean section (N) • No: 679 • Yes: 90 Non-PPD population N = 9,211 Age (mean): 33.92 years Marital status (N) • Single: 1,470 • Married: 4,610 • Not known: 3,131 Caesarean section (N) • No: 8,352 • Yes: 859 **ML prediction model performance for PPD** SVM model vs. ICD-10 codes indicating PPD post-birth AUC: 0.79 SEN: 0.894 SPE: 0.580	"In this pilot study, we demonstrate promising PPD prediction results using a machine learning approach with information on patient demographics, diagnoses, and medications available from EHRs. Our goal is to create an accurate PPD prediction model to identify risk factors for PPD and facilitate effective screening of mothers who may require early intervention for PPD using an EHR. We envision that the model may be integrated with the EHR system for a provider-facing CDS or with a mobile or web platform to be used as a patient-facing CDS in a future phase of the study" (p. 891).

Main Study Findings	Authors' Conclusion
L2 LR Model vs. ICD-10 codes indicating PPD post-birth AUC: 0.78 SEN: 0.887 SPE: 0.594 RF Model vs. ICD-10 codes indicating PPD post-birth AUC: 0.78 SEN: 0.959 SPE: 0.391 Naive Bayes model vs. ICD-10 codes indicating PPD post-birth AUC: 0.78 SEN: 0.867 SPE: 0.616 XGBoost model vs. ICD-10 codes indicating PPD post-birth AUC: 0.77 SEN: 0.915 SPE: 0.527 Decision tree model vs. ICD-10 codes indicating PPD post-birth AUC: 0.69 SEN: 0.986 SPE: 0.386	

Main Study Findings	Authors' Conclusion
Prediction results in different variable combination using SVM (best performing model) Predictors Trimesters 1st: 0.66 AUC; 0.855 SEN; 0.428 SPE 2nd: 0.64 AUC; 0.831 SEN; 0.424 SPE 3rd: 0.65 AUC; 0.867 SEN; 0.424 SPE 1st + 2nd: 0.69 AUC; 0.908 SEN; 0.307 SPE 2nd + 3rd: 0.72 AUC; 0.854 SEN; 0.524 SPE Categories Demographic: 0.60 AUC; 0.551 SEN; 0.609 SPE Diagnose: 0.72 AUC; 0.850 SEN; 0.560 SPE Medication: 0.65 AUC; 0.882 SEN; 0.389 SPE Diagnose + medication: 0.76 AUC; 0.875 SEN; 0.577 SPE Logistic-selected: 0.76 AUC; 0.892 SEN; 0.588 SPE	

Main Study Findings	Authors' Conclusion
Zhao 2019[49]	
Demographics N = 179 (100 males, 79 females) Average Age: 24.2 **Predictive accuracies of the regression models on GAD-7 scores** SLR vs. GAD-7 and PHQ-9: −0.07 LR vs. GAD-7 and PHQ-9: 0.24 ($P < 0.01$) e-SVR vs. GAD-7 and PHQ-9: 0.51 ($P < 0.001$) n-SVR vs. GAD-7 and PHQ-9: 0.48 ($P < 0.001$) GP vs. GAD-7 and PHQ-9: 0.43 ($P < 0.001$) **Predictive accuracies of the regression models on PHQ-9 scores** SLR vs. GAD-7 and PHQ-9: -0.16 ($P < 0.05$) LR vs. GAD-7 and PHQ-9: 0.23 ($P < 0.01$) e-SVR vs. GAD-7 and PHQ-9: 0.38 ($P < 0.001$) n-SVR vs. GAD-7 and PHQ-9: 0.40 ($P < 0.001$) GP vs. GAD-7 and PHQ-9: 0.51 ($P < 0.001$)	"Our experiment demonstrated that the natural gaits could be an objective data source for measuring anxiety and depression, and the predictive models showed the effectiveness not only in recognizing the total questionnaire scores of anxiety and depression, but also in detecting some self-reported specific depressive symptoms. Though the nonpatient sample and the questionnaire-based design limited the applicability of the current model, this pilot study indicated one possible direction that is worthy of further investigation for new convenient mental health measuring methods" (p. 11).
Zhuang (2019)[51]	
Demographics N = 69 (40 drug-naive FES patients; 29 HC) Age: FES patients = 27.13; HC = 27.03 Gender (females/males): FES patients = 18/22; HC = 15/14 Education (years): FES patients = 12.91; HC = 14.21 **Comparison of classification performance for FES patients** SVM model only vs. clinician assessment (*DSM-IV*) sMRI feature: 61.43% accuracy; 85% SEN; 30% SPE; 54.66% AUC DTI feature: 60.48% accuracy; 65% SEN; 53.33% SPE; 50.60% AUC FC feature: 74.05% accuracy; 82.50% SEN; 63.33% SPE; 74.48% AUC	"In this paper, we proposed a multimodal classification framework using multi-kernel and sparse coding machine learning method for drug naïve FES patients diagnosis. We combined multimodal MR imaging data, including structural MR images, diffusion tension images and resting-state functional MR images. To effectively improve the performance of the SVM classifier, we applied sparse coding to reduce the feature dimension and identify the most discriminative image biomarkers. The best classification performance was achieved when incorporating all anatomical, diffusion and functional image data for the subjects" (p. 93).

Main Study Findings	Authors' Conclusion
fALFF feature: 60.71% accuracy; 80% SEN; 35% SPE; 56.29% AUC Multimodal: 76.67% accuracy; 95% SEN; 58.33% SPE; 42.84% AUC SC + SVM model vs. clinician assessment (*DSM-IV*) sMRI feature: 71.19% accuracy; 77.50% SEN; 63.33% SPE; 69.19% AUC DTI feature: 67.86% accuracy; 72.50% SEN; 61.67% SPE; 68.19% AUC FC feature: 75.24% accuracy; 80% SEN; 68.33% SPE; 75.26% AUC fALFF feature: 69.29% accuracy; 80% SEN; 55% SPE; 61.90% AUC Multimodal: 84.29% accuracy; 92.50% SEN; 73.33% SPE; 81.64% AUC	
Fulmer (2018)[33]	
Demographics Control group: n = 24 67% female (4% nonconforming) Mean age = 22.5 46% white 33% Asian 13% other 8% AA Group 1 (Tess for 2 weeks): n = 24 71% female Mean age = 22.2 54% white 46% Asian Group 2 (Tess for 4 weeks): n = 26 73% female Mean age = 22 69% Asian 31% white	"Results revealed that both test groups 1 and 2 experiences a significant reduction in symptoms of anxiety with unlimited access to Tess over the course of 2 or 4 weeks. Furthermore, the test group that received daily check-ins from Tess over 2 weeks experiences a significant reduction in symptoms of depression. Participants who interested with Tess displayed higher levels of engagement and overall satisfaction than those from the control group. Test group participants indicated that the content was more relevant to their everyday life and made them more comfortable with the therapeutic experience" P. 9.

Main Study Findings	Authors' Conclusion
Outcomes Control group, mean score (baseline) PHQ-9 = 8.17 GAD-7 = 9.46 PANAS positive affect = 22.13 PANAS negative affect = 15.75 Group 1, mean score (baseline) PHQ-9 = 6.67 GAD-7 = 6.71 PANAS positive affect = 19.88 PANAS negative affect = 13.08 Group 2, mean score (baseline) PHQ-9 = 7.04 GAD-7 = 7.5 PANAS positive affect = 21.31 PANAS negative affect = 14.38 PHQ-9 Control vs. group 1; *P* = 0.02 GAD-7 Control vs. group 1, significant; *P* = NR Control vs. group 2, significant; *P* = NR Group 1, change from baseline; *P* = 0.045 Group 2, change from baseline; *P* = 0.2 Control, change from baseline; *P* = NS	

Main Study Findings	Authors' Conclusion
PANAS Control vs. group 1; *P* = 0.03 User satisfaction and engagement Satisfaction Significant difference between control and test groups; *P* = NR Control group = 60% satisfaction Test groups = 86%	
McGinnis (2018)[39]	
Demographics N = 63 57% female 65% white **Classification accuracy of different models for each feature set** Accuracy DT model vs. multimodal assessments (diagnostic interviews) ACC = 58% GYR= 69% ANG = 68% ACC + ANG = 63% ACC + GYR = 66% GYR + ANG = 69% ACC + GYR + ANG = 69% KNN vs. multimodal assessments (diagnostic interviews) ACC = 53% GYR = 59%	"The results presented herein demonstrate that, when paired with ML, 20 seconds of wearable sensor data extracted from a fear induction task can be used to diagnosis internalizing disorder in young children with a high level of accuracy and at a fraction of the cost and time of existing assessment techniques" (p. 3986).

Main Study Findings	Authors' Conclusion
ANG = 71% ACC + ANG = 56% ACC + GYR = 73% GYR + ANG = 69% ACC + GYR + ANG = 73% SVM vs. multimodal assessments (diagnostic interviews) ACC = 64% GYR = 71% ANG = 78% ACC + ANG = 64% ACC + GYR = 76% GYR + ANG = 80% ACC + GYR + ANG = 76% LR vs. multimodal assessments (diagnostic interviews) ACC = 64% GYR = 71% ANG = 78% ACC + ANG = 66% ACC + GYR = 75% GYR + ANG = 76% ACC + GYR + ANG = 80% SVM with GYR + ANG and LR with ACC + GYR + ANG had the best accuracy (80%) ROC curves SVM = 0.92 LR = 0.89	

Main Study Findings	Authors' Conclusion
He (2017)[34]	
Demographics N = 300 trauma survivors (150 diagnosed as PTSD patients and 150 as non-PTSD patients) **Comparison between ML text classifiers with N-grams** DT accuracy vs. diagnosis obtained by the practitioners via structured interviews with standardized instruments (*DSM-IV* and Clinician-Administered PTSD Scale) Unigrams: 0.57 Bigrams: 0.60 Trigrams: 0.57 Unigrams + Bigrams: 0.58 Unigrams + Bigrams + Trigrams: 0.58 NB accuracy vs. diagnosis obtained by the practitioners via structured interviews with standardized instruments (*DSM-IV* and Clinician-Administered PTSD Scale) Unigrams: 0.79 Bigrams: 0.68 Trigrams: 0.60 Unigrams + Bigrams: 0.78 Unigrams + Bigrams + Trigrams: 0.76 SVM accuracy vs. diagnosis obtained by the practitioners via structured interviews with standardized instruments (*DSM-IV* and Clinician-Administered PTSD Scale) Unigrams: 0.80 Bigrams: 0.57 Trigrams: 0.53 Unigrams + Bigrams: 0.70 Unigrams + Bigrams + Trigrams: 0.69	"The results showed that the textual assessment on self-narratives achieved a high agreement with practitioners' diagnoses, and the addition of higher order n-grams could help balance the classification metrics and enhance the reliability of classification prediction. This article further demonstrates that the automated textual assessment system is a promising tool for analyzing patients' selfexpression behaviors, thus helping practitioners identify potential patients at an early stage" (p. 170).

Main Study Findings	Authors' Conclusion
PSM Accuracy vs. diagnosis obtained by the practitioners via structured interviews with standardized instruments (*DSM-IV* and Clinician-Administered PTSD Scale) Unigrams: 0.82 Bigrams: 0.76 Trigrams: 0.67 Unigrams + Bigrams: 0.81 Unigrams + Bigrams + Trigrams: 0.80 **Comparison between ML text classifiers with unigrams by different prevalence of PTSD (Accuracy)** 5% prevalence of PTSD DT: 0.57 NB: 0.78 SVM: 0.76 PSM: 0.80 15% prevalence of PTSD DT: 0.56 NB: 0.78 SVM: 0.76 PSM: 0.79 25% prevalence of PTSD DT: 0.57 NB: 0.77 SVM: 0.78 PSM: 0.80	

Main Study Findings	Authors' Conclusion
50% prevalence of PTSD DT: 0.57 NB: 0.79 SVM: 0.80 PSM: 0.82 Mean performance: 0.86	
Wang (2017)[47]	
Accuracy of CMS-HCC R^2 = 0.09 PCA = 27%, top 10% high-cost patient cut-off CA = 45%, top 10% high-cost patient cut-off PCA = 35%, top 20% high-cost patient cut-off CA = 57%, top 20% high-cost patient cut-off **Accuracy of baseline model** R^2 = 0.19 PCA = 40%, top 10% high-cost patient cut-off CA = 58%, top 10% high-cost patient cut-off PCA = 50%, top 20% high-cost patient cut-off CA = 66%, top 20% high-cost patient cut-off **Accuracy of enhanced model** R^2 = 0.24 PCA = 42%, top 10% high-cost patient cut-off CA = 61%, top 10% high-cost patient cut-off PCA = 52%, top 20% high-cost patient cut-off CA = 68%, top 20% high-cost patient cut-off	"We found, using advanced feature selection and supervised machine learning methods, and leveraging detailed clinical and medication data, that there was an improvement in the ability to predict and identify high-cost patients with SCZ compared with the CMS-HCC model. Improving our ability to predict high-cost/high-risk patients with mental health issues including SCZ may provide support to health organizations to coordinate and deliver the right services to the most appropriate individuals" (p. 6).

Main Study Findings	Authors' Conclusion
Accuracy of final model R^2 = 0.24 PCA = 43%, top 10% high-cost patient cut-off CA = 63%, top 10% high-cost patient cut-off PCA = 53%, top 20% high-cost patient cut-off CA = 69%, top 20% high-cost patient cut-off	
Devine (2016)[31]	
Demographics Age, mean 44.7 Gender, % female 69.1 Main diagnoses (ICD-10), % Somatoform disorder (F45.x) = 23.4 Depressive disorder (F3x.x) = 20.0 Dissociative disorder (F44.x) = 15.6 Eating disorders (F50.x) = 10.3 Anxiety disorders (F40/1.x) = 6.2 Adjustment disorders (F43.x) = 1.4 Other mental disorders = 3 Other medical conditions = 20.1 **Outcomes** Measurement precision (SE) CAT = 0.30; 95% CI, 0.59 on z metric Theta score range = 6 to 9 SD (measurement precision set at 0.32 or less)	"To conclude, the tested German mental health CATs are the first CATs world wide being used in daily routines for more than a decade. At that time state-of-the art methods have been used, which are still applied e.g. in the PROMIS project. Our results show the benefits of CAT assessment for monitoring mental health: They are short, efficient and comparable in response burden to existing static short forms. Retest-reliability is comparable to established tools, while sensitivity to change seems similar to lower – though not significantly lower than of traditional measures. Potential explanations for this need to be elucidated in future longitudinal mental health CAT studies… Also it is advisable to include clinical variables in order to establish minimal clinically important changes scores" (p. 851).

Main Study Findings	Authors' Conclusion
Number of items administered, average D-CAT = 5.6 A-CAT = 5.7 S-CAT = 7.2 Retest-reliability rD-CAT= 0.71 vs. rPHQ-9 = 0.75 rA-CAT= 0.78 vs. rGAD-7 = 0.75 rS-CAT = 0.80 vs. rPSQ"worries" = 0.80 Sensitivity to change (between admission and discharge) "All mean scores decreased between admission and discharge, thus all questionnaires were able to capture the average significant improvement of the patients over time" P. 850. All $P < 0.001$ Cohen's D 95% CI D-CAT = 0.25 to 0.81 PHQ-9 = 0.40 to 0.98 A-CAT = 0.01 to 0.56 GAD-7 = 0.21 to 0.77 S-CAT = −0.09 to 0.47 PSQ = −0.11 to 0.45 Note: All CATs were compared with conventional depression, stress, and anxiety instruments (PHQ-9, GAD-7, and PSQ).	

Main Study Findings	Authors' Conclusion
Achtyes (2015)[10]	
Demographics N = 145 79% were female MDD: n = 27 GAD: n = 27 BPD1: n = 13 BPD2: n = 11 Dysthymic disorder: n = 15 Minor depression: n = 2 Panic disorder: n = 16 Agoraphobia : n = 6 Social phobia : n = 13 Specific phobia: n = 9 OCD: n = 11 PTSD: n = 12 Anxiety, not otherwise specified: n = 15 **Outcomes** CAT-MDD Sensitivity = 0.96 ("0.95 in the original CAD-MDD study" P. 6) Specificity = 0.64 ("0.87 in the original CAD-MDD study which included a much greater number and proportion of controls" P. 6). Restricted sample (only patient with *DSM-IV* criteria for MDD) Sensitivity = 0.96 Specificity = 1.00 Average of 4.1 questions Average time of 36.1 seconds	"The results of this prospective, cross-sectional validation study suggest that the CAT-MH suite of tests provide a rapidly-administered, accurate assessment of depression diagnosis and symptom severity across a broad range of mood and anxiety symptoms in an adult, community outpatient psychiatric population" (p. 8).

Main Study Findings	Authors' Conclusion
CAT-DI Depression severity correlations HAM-D_{25}, (r = 0.79), PHQ-9 (r = 0.90), CES-D (r = 0.90), and GAF (r = -0.70), CAT-ANX (r = 0.82), CAT-MANIA (r = 0.38) With current *DSM-IV* diagnosis, OR = 6.97 (95% CI, 3.14 to 15.51); *P* < 0.001 Average of 16.8 questions Average time of 3.4 mins CAT-ANX Anxiety severity correlations HAM-D25 (r = 0.73), PHQ-9 (r = 0.78), CES-D (r = 0.81), and GAF (r = −0.68), CAT-MANIA (r = 0.47) OR = 2.88 (95% CI, 1.72 to 4.83); P < 0.001 Average of 12.9 questions Average time of 2.0 mins CAT-MANIA Correlations HAM-D_{25} (r = 0.31), PHQ-9 (r = 0.37), CES-D (r = 0.39), and GAF (r = −0.29) OR = 2.89 (95% CI, 1.47 to 5.71); *P* < 0.002 Average of 17.9 questions Average time of 3.4 mins Note: All CATs were compared with conventional depression, stress, and anxiety instruments.	

Main Study Findings	Authors' Conclusion
Jimenez-Serrano (2015)[36]	
Demographics All participants white No participants were in psychiatric treatment during the pregnancy **Classification Results** Tortajada et al. (comparator: hold-out evaluation): NOV = 16 NOW = 8-32 G = 0.81 SEN = 0.78 SPE = 0.85 AUC = 0.84 ACC = 0.84 Naive Bayes vs. Spanish EPDS test version and DIGS: NOV = 11 NOW = 1 G = 0.73 SEN = 0.72 SPE = 0.73 AUC = 0.75 ACC = 0.73	"Different models for predicting PPD have been developed using ML and PR techniques. These models have the ability to predict PPD during the first week after childbirth with a reasonable accuracy. Finally, the model that achieved the best balance between sensitivity and specificity was integrated into a CDSS for Android mobile apps. This approach can enable the early prediction and detection of PPD because it fulfills the conditions of an effective test with an acceptable level of sensitivity and specificity that is quick to perform, easy to interpret, culturally sensitive, and cost-effective. The mobile app can be clinically evaluated in future works" (p. 573). "Among all the trained models during the experimentation, the naive Bayes model presented the best performance on the test dataset according to G function, with a value of 0.73. A good balance among sensitivity, specificity, and accuracy was achieved, with values close to 0.73 in all cases. Thus, a new naive Bayes model with the above best hyperparameters and using all the available data from 11 independent variables was retrained and integrated into the Android mobile app" (p. 572).

Main Study Findings	Authors' Conclusion
Logistic regression vs. Spanish EPDS test version and DIGS: NOV = 11 NOW = 1 G = 0.69 SEN = 0.63 SPE = 0.75, AUC = 0.7 ACC = 0.74 SVM vs. Spanish EPDS test version and DIGS: NOV = 11 NOW = 1 G = 0.65 SEN = 0.56 SPE= 0.75 AUC = 0.75 ACC = 0.73 ANN vs. Spanish EPDS test version and DIGS: NOV=11 NOW=1 G=0.6 SEN = 0.53 SPE=0.83 AUC=0.66 ACC=0.79	

Main Study Findings	Authors' Conclusion
Zhou (2015)[50]	
Demographics N = 1200 All patients with history of heart disease Number of depression cases, reference test (i.e., "gold standard") Training data set High confidence: n = 89 Intermediate confidence: n = 22 (Note: High confidence = when depression diagnosis terms present in notes; intermediate confidence = combinations of antidepressant treatment, psychiatry consultation, or depressive symptomatology present in notes) Testing data set High confidence: n = 79 Intermediate confidence: n = 31 **Performance of NLP models** High confidence (n = 79) MTERMS vs. manual review by domain experts: Precision = 86.9% Recall = 92.4% F-measure = 89.6% C4.5 vs. manual review by domain experts: Precision = 87.8% Recall = 91.1% F-measure = 89.4%	"Our system achieved an F-measure of 89.6% in identifying high confidence cases and 70.6% for intermediate confidence cases. MTERMS' performance was slightly better than machine learning classifiers. Recall was higher than precision (92.4% vs. 86.9% for high confidence cases and 77.4% vs. 64.9% for intermediate confidence cases), which indicates that such a system will be useful for retrieving relevant instances. Identified cases can then be reviewed by clinicians and researchers. By examining the classification errors made by the system, we found that there were several cases in which our knowledge-based system classified a patient who did not have depression as depressed with either high or intermediate confidence. In some of these false positive cases, a depression-related term was negated, but the negation was outside our NLP algorithm's scope… Our lexicon was able to correctly discover some but not all of these phrases. In addition, there were a few cases in which a patient possibly having depression with intermediate confidence was not identified by our system. These false-negative cases occurred mainly because the system did not identify symptoms outside our lexicon" (p. 632).

Main Study Findings	Authors' Conclusion
NNge vs. manual review by domain experts: Precision = 87.8% Recall = 91.1% F-measure = 89.4% RIPPER vs. manual review by domain experts: Precision = 85.7% Recall = 91.1% F-measure = 88.3% SVM vs. manual review by domain experts: Precision = 86.7% Recall = 91.1% F-measure = 88.9% Intermediate confidence (n = 31) MTERMS decision tree vs. manual review by domain experts: Precision = 64.9% Recall = 77.4% F-measure = 70.6% C4.5 vs. manual review by domain experts: Precision = 64.0% Recall = 51.6% F-measure = 57.1%	

Main Study Findings	Authors' Conclusion
NNge vs. manual review by domain experts: Precision = 75.0% Recall = 29.0% F-measure = 41.9% RIPPER vs. manual review by domain experts: Precision = 72.4% Recall = 67.7% F-measure = 70.0% SVM vs. manual review by domain experts: Precision = 65.2% Recall = 48.4% F-measure = 55.6%	

AF = air force; 1NN = nearest neighbour; 3NN = three-nearest neighbour; AA = African American; ACC = accelerometer feature; ACC = accuracy; ADHD = attention-deficit/hyperactivity disorder; AMA = anticorrelation after mean of antilog; ANG = angle feature; ANN = artificial neural network; ANS = autonomic nervous system; AUC = area under the curve; BPD = bipolar disorder; BRI = barrier reduction intervention; CA = cost accuracy; CAD-MDD = Computerized Adaptive Diagnostic Test for Major Depressive Disorder; CAE = 3-D convolutional autoencoder; CAT = computerized adaptive test; CAT-ANX = computerized adaptive test for anxiety severity; CAT-DI = computerized adaptive test for depression severity; CAT-MANIA = computerized adaptive test for manic/hypomanic symptom severity; CAT-MH = computerized adaptive test for mental health; CDS = Clinical decision support; CDSS = clinical depression support system; CES-D = Center for Epidemiologic Studies Depression Scale; CI = confidence interval; CMS-HCC = Centers for Medicare & Medicaid Services Hierarchical Condition Categories; CNN = convolutional neural network; DL = deep learning; DLDA = diagonal linear discriminate analysis; DIGS = Diagnostic interview for genetic studies; *DSM-IV = Diagnostics and Statistics Manual of Mental Disorders, Fourth Edition*; DT = decision tree; DTI = diffusion tension imaging; EEG = electroencephalogram; EHR = electronic health record; EPDS = Edinburgh Postnatal Depression Scale; e-SVR = Epsilon-SVR; FC = functional connectivity; FES = first-episode schizophrenia; fMRI = functional magnetic resonance imaging; GAD = general anxiety disorder; GAD-7 = Generalized Anxiety Disorder 7-item scale; GAF = Global Assessment of Functioning; GM = grey matter; GP = gaussian process; GSR = galvanic skin response; GYR = gyro feature; HAMD-D25 = Hamilton Rating Scale for Depression; HC = health control; HRV = hear rate variability; ICD-10 = International Statistical Classification of Diseases and Related Health Problems 10; ITT = intent to treat; K-NHANES = Korea–National Health and Nutrition Examination Survey; KNN = k-nearest neighbour; L2 LR = L2-regularized logistic regression; LDA = linear discriminate analysis; LR = logistic regression; MCC = Matthews coefficient correlation; MDD = major depressive disorder; ML = machine learning; MR = magnetic resonance; MTERMS = medical text extraction reasoning and mapping system; NB = naive Bayes; NC = nearest centroid; NF = number of features; NHANES = National Health and Nutrition Examination Survey; NLP = natural language processing; NN = nearest neighbor; NNge = generalized nearest neighbour; NNT = number needed to treat; NOV = number of variables; NOW = number of weeks; NPV = negative predictive value; NR = not reported; NS = not specified; n-SVR = Nu-SVR; OCD = obsessive compulsive disorder; OR = odds ratio; PANAS = Positive and Negative Affect Schedule; PCA = principal component analysis; PCL-C = PTSD Checklist Civilian Version; PHQ-9 = Patient Health Questionnaire-9; PPD = postpartum depression; PPV = positive predictive value; PR = pattern recognition; PROMIS = Patient-Reported Outcome Measurement Information System; PSM = product score model; PSQ = Perceived Stress Questionnaire; PTSD = post-traumatic stress disorder; RF = random forest; RIPPER = Repeated Incremental Pruning to Produce Error Reduction; ROC = receiver operating characteristic; RR = relative risk; RRI = R-peaks in the EEG signal; SC = sparse coding; SCZ = schizophrenia; SD = standard deviation; SDS = Self-Rating Depression Scale; SE = sensitivity; SEN = sensitivity; SG = support vector machine with gaussian kernel; SL = support vector machine with linear kernel; SLR = simple linear regression; sMRI = structural magnetic resonance imaging; SPE = specificity; SSD = schizophrenic spectrum disorder; SVM = support vector machine; SVM-PCA = support vector machine with principal component analysis; SVM-RFE = support vector machine learning with recursive feature elimination; TE = trauma exposed; UK AF = United Kingdom Armed Forces; VBM = voxel-based morphometry.

Appendix 6: Overlap Between Included Systematic Reviews

Table 9: Primary Study Overlap Between Included Systematic Reviews

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Acikel et al. (2016)							X	
Akinci et al. (2013)							X	
Al-Kaysi et al. (2017)						X		
Almeida et al. (2013)							X	
Amin et al. (2018)			X					
Ammerman et al. (2017a)	X							
Ammerman et al. (2017b)	X							
Ammerman et al. (2018)	X							
Anticevic et al. (2014)							X	
Arbabshirani (2014)			X					
Arribas et al. (2010)							X	
Baca-Garcia et al. (2010)	X							
Bae et al. 2015	X							
Bae et al. (2017)			X					
Bailey et al. (2018)						X		
Barros et al. (2017)	X							
Batterham and Christensen (2012)	X							
Belzeaux et al. (2016)						X		
Besga et al. (2012)							X	
Besga et al. (2015)							X	
Bhaumik et al. (2017)				X				
Bollen et al. (2011)								X
Braithwaite et al. (2016)	X							X
Burger et al. (2017)				X				
Burke et al. (2018)	X							
Burnap et al. (2015)								X
Burns, M.N. (2011)		X						
Cabral et al. (2016)			X					

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Cao et al. (2014)			X	X				
Castellani et al. (2012)			X					
Castro et al. (2014)			X					
Chancellor et al. (2016)								X
Chekroud et al. (2016)						X		
Chekroud et al. (2017)						X		
Chen et al. (2014)							X	
Chen H et al. (2017)			X					
Chen X et al. (2017)			X					
Cheng et al. (2017)	X							
Chuang and Kuo (2017)							X	
Chyzhyk et al. (2015)			X					
Cook et al. (2016)	X							
Coppersmith et al. (2014a)								X
Coppersmith et al. (2014b)								X
Coppersmith et al. (2015)								X
Coppersmith et al. (2016)								X
Costafreda et al. (2009)				X		X		
Costafreda et al. (2011)							X	
De Choudhury et al. (2013a)								X
De Choudhury et al. (2013b)								X
De Choudhury et al. (2013c)								X
De Choudhury et al. (2014)								X
Delgado-Gomez et al. (2016)	X							
Deng et al. (2017)				X				
Dmitrzak-Weglarz et al. (2014)							X	
Drysdale et al. (2016)				X				
Drysdale et al. (2017)						X	X	
Du et al. (2015)							X	
Durahim et al. (2015)								X
Erguzel et al. (2015)							X	
Erguzel et al. (2016)							X	

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Etkin et al. (2015)						X		
Fang et al. (2012)				X				
Fitzpatrick et al. (2017)					X			
Foland-Ross et al. (2015)				X				
Frangou et al. (2016)				X				
Frangou et al. (2017)							X	
Fu et al. (2008)				X				
Fung et al. (2015)				X			X	
Gao et al. (2017)				X				
Gong et al. (2011)				X				
Gradus et al. (2017)	X							
Greenstein (2012)			X					
Grotegard et al. (2013b)				X				
Grotegerd et al. (2013)				X			X	
Grotegerd et al. (2014)							X	
Guan et al. (2015)	X							X
Guilloux et al. (2015)						X		
Guo et al. (2014)				X				
Guo (2017)			X					
Habes et al. (2013)				X				
Haenisch et al. (2016)							X	
Hajek et al. (2015)							X	
Handley et al. (2014)	X							
Hao et al. (2013)								X
Hao et al. (2014)								X
He et al. (2017)				X				
Hettige et al. (2017)	X							
Hilbert et al. (2017)				X				
Hill et al. (2017)	X							
Homan et al. (2014)								X
Hu et al. (2015)								X
Huang et al. (2014)								X

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Hudlicka (2013)					X			
Ilgen et al. (2009)	X							
Iniesta et al. (2016)						X		
Iwabuchi (2013)			X					
Jain et al. (2013)						X		
Jamison-Powell et al. (2012)								X
Jie et al. (2015)				X				
Jie et al. (2015)				X			X	
Jing et al. (2017)				X				
Johannesen et al. (2012)							X	
Johnston et al. (2015a)				X				
Johnston et al. (2015b)				X				
Jordan et al. (2018)	X							
Just et al. (2017)	X							
Kang et al. (2016)								X
Kaufmann et al. (2017)							X	
Kautzky et al. (2015)						X		
Kautzky et al. (2017)						X		
Kessler et al. (2015)	X							
Kessler et al. (2017a)	X							
Kessler et al. (2017b)	X							
Khodayari-Rostamabad et al. (2010)							X	
Khodayari-Rostamabad et al. (2011)						X		
Khodayari-Rostamabad et al. (2013)						X		
Kim et al. (2016)			X					
Koch et al. (2015)			X					
Korgaonkar et al. (2014)				X				
Korgaonkar et al. (2015)				X		X		
Koutsouleris et al. (2015)				X			X	
Kuang et al. (2014)								X
Kuroki and Tilley (2012)	X							
Kuroki et al. (2015)	X							

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Landeiro Dos Reis et al. (2015)								X
Li et al. (2017)				X				
Lin et al. (2014)								X
Liu et al. (2015)								X
Liu et al. (2012)				X		X		
Liu (2017)			X					
Liu (2018)			X					
Lopez-Castroman et al. (2011)	X							
Lord et al. (2012)				X				
Lu et al. (2016)			X					
Lucas et al. (2017)					X			
Lv et al. (2015)								X
Lythe et al. (2015)				X				
Ma et al. (2013)				X				
Macmaster et al. (2014)				X				
Mann et al. (2008)	X							
Marquand et al. (2008)						X		
Matsuraba (2015)			X					
Metzger et al. (2017)	X							
Miner et al. (2016)					X			
Modinos et al. (2013)				X				
Mohr et al. (2017)		X						
Morales et al. (2017)	X							
Mourão-Miranda et al. (2011)				X				
Mourão-Miranda et al. (2012)							X	
Mumtaz et al. (2017)						X		
Mwangi et al. (2012)				X				
Mwangi et al. (2016)							X	
Nouretdinov et al. (2011)				X				
Oh et al. (2017)	X							
Orban (2017)			X					
Park et al. (2013)								X

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Park et al. (2015)								X
Passos et al. (2016)	X							
Patel et al. (2015)				X		X		
Pederson et al. (2015)								X
Pestian et al. (2016)	X							
Pestian et al. (2017)	X							
Philip et al. (2017)					X			
Pinaya (2016)			X					
Pinaya (2019)			X					
Pinto et al. (2017)							X	
Pirooznia et al. (2012)							X	
Pläschke et al. (2017)			X					
Poulin et al. (2014)	X							
Preotiuc-Pietro et al. (2015)								X
Prieto et al. (2014)								X
Qureshi (2017)			X					
Ramasubbu et al. (2016)				X				
Reavis et al. (2017)			X					
Redlich et al. (2014)				X			X	
Redlich et al. (2016)						X		
Resnik et al. (2015a)								X
Resnik et al. (2015b)								X
Rive et al. (2016)				X			X	
Roberts et al. (2016)							X	
Rocha-Rego et al. (2014)							X	
Rondina et al. (2014)				X				
Rosa et al. (2015)				X				
Rubin-Falcone et al. (2017)				X				
Sacchet et al. (2015b)				X				
Sacchet et al. (2015)				X			X	
Salvador (2017)			X					
Sankar et al. (2016)				X				

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Saravia et al. (2016)								X
Sato et al. (2015)				X				
Schmaal et al. (2015)				X				
Schnack et al. (2014)							X	
Schnyer et al. (2017)				X				
Schwartz et al. (2013)								X
Schwartz et al. (2014)								X
Schwartz et al. (2016)								X
Serpa et al. (2014)				X			X	
Serretti and Smeraldi (2004)						X		
Serretti et al. (2007)						X		
Shimizu et al. (2015)				X				
Struyf et al. (2008)							X	
Su (2013)			X					
Sundermann et al. (2017)				X				
Tanaka et al. (2017)					X			
Thibodeau et al. (2015)						X		
Tsugawa et al. (2013)								X
Tsugawa et al. (2015)								X
van Waarde et al. (2014)					X	X		
Volkava et al. (2016)								X
Wade et al. (2016)						X		
Walsh et al. (2017)	X							
Wang et al. (2013a)								X
Wang et al. (2013b)								X
Wang et al. (2017c)				X				
Wang et al. (2017d)								X
Wang et al. (2017a)			X					
Wang (2017b)			X					
Watanabe et al. (2014)			X					
Wei et al. (2013)				X				
Williams et al. (2015)				X				

Primary Study Citation	Systematic Review Citation							
	Burke (2019)[21]	Colombo (2019)[20]	de Filippis (2019)[24]	Gao (2018)[7]	Laranjo (2018)[13]	Lee (2018)[22]	Passos (2019)[23]	Wongkoblap (2019)[25]
Wilson et al. (2014)								X
Wu et al. (2016a)							X	
Xiao et al. (2017)			X					
Yang et al. (2016)				X				
Yang et al. (2010)			X					
Yoon et al. (2008)			X					
Yoshida et al. (2017)				X				
Yu et al. (2013)				X				
Zeng et al. (2012)				X				
Zeng et al. (2014)				X				
Zeng et al. (2018)			X					
Zhang et al. (2015)								X
Zhong et al. (2017)				X				