Current Psychiatry Reports (2019) 21:116
https://doi.org/10.1007/s11920-019-1094-0

PSYCHIATRY IN THE DIGITAL AGE (J SHORE, SECTION EDITOR)

Artificial Intelligence for Mental Health and Mental Illnesses: an Overview

Sarah Graham[1,2] · Colin Depp[1,2,3] · Ellen E. Lee[1,2,3] · Camille Nebeker[4] · Xin Tu[1,2] · Ho-Cheol Kim[5] · Dilip V. Jeste[1,2,6,7]

Abstract

Purpose of Review Artificial intelligence (AI) technology holds both great promise to transform mental healthcare and potential pitfalls. This article provides an overview of AI and current applications in healthcare, a review of recent original research on AI specific to mental health, and a discussion of how AI can supplement clinical practice while considering its current limitations, areas needing additional research, and ethical implications regarding AI technology.

Recent Findings We reviewed 28 studies of AI and mental health that used electronic health records (EHRs), mood rating scales, brain imaging data, novel monitoring systems (e.g., smartphone, video), and social media platforms to predict, classify, or subgroup mental health illnesses including depression, schizophrenia or other psychiatric illnesses, and suicide ideation and attempts. Collectively, these studies revealed high accuracies and provided excellent examples of AI's potential in mental healthcare, but most should be considered early proof-of-concept works demonstrating the potential of using machine learning (ML) algorithms to address mental health questions, and which types of algorithms yield the best performance.

Summary As AI techniques continue to be refined and improved, it will be possible to help mental health practitioners re-define mental illnesses more objectively than currently done in the DSM-5, identify these illnesses at an earlier or prodromal stage when interventions may be more effective, and personalize treatments based on an individual's unique characteristics. However, caution is necessary in order to avoid over-interpreting preliminary results, and more work is required to bridge the gap between AI in mental health research and clinical care.

Keywords Technology · Machine learning · Natural language processing · Deep learning · Schizophrenia · Depression · Suicide · Bioethics · Research ethics

This article is part of the Topical Collection on *Psychiatry in the Digital Age*

✉ Dilip V. Jeste
djeste@ucsd.edu

[1] Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

[2] Sam and Rose Stein Institute for Research on Aging, University of California La Jolla, La Jolla, CA, USA

[3] VA San Diego Healthcare System, San Diego, CA, USA

[4] Department of Family Medicine and Public Health, University of California La Jolla, La Jolla, CA, USA

[5] Scalable Knowledge Intelligence, IBM Research-Almaden, San Jose, CA, USA

[6] Department of Neurosciences, University of California La Jolla, La Jolla, CA, USA

[7] University of California San Diego, 9500 Gilman Drive, Mail Code #0664, La Jolla, CA 92093-0664, USA

Introduction and Background of Artificial Intelligence in Healthcare

We are at a critical point in the fourth industrial age (following the mechanical, electrical, and internet) known as the "digital revolution" characterized by a fusion of technology types [1, 2]. A leading example is a form of technology originally recognized in 1956—artificial intelligence (AI) [3]. While several prominent sectors of society are ready to embrace the potential of AI, caution remains prevalent in medicine, including psychiatry, evidenced by recent headlines in the news media like "A.I. Can Be a Boon to Medicine That Could Easily Go Rogue" [4]. Regardless of apparent concerns, AI applications in medicine are steadily increasing. As mental health practitioners, we need to familiarize ourselves with AI, understand its current and future uses, and be prepared to knowledgeably work with AI as it enters the clinical mainstream [5]. This article provides an overview of AI in healthcare (introduction), a review of

Published online: 07 November 2019

original, recent literature on AI and mental healthcare (methods/results), and a discussion of how AI can supplement mental health clinical practice while considering its current limitations, identification of areas in need of additional research, and ethical implications (discussion/future directions).

AI in Our Daily Lives

The term AI was originally coined by a computer scientist, John McCarthy, who defined it as "the science and engineering of making intelligent machines" [6]. Alan Turing, considered to be another "father of AI," authored a 1950 article, "Computing Machinery and Intelligence" that discussed conditions for considering a machine to be intelligent [7]. As intelligence is traditionally thought of as a human trait, the modifier "artificial" conveys that this form of intelligence describes a computer. AI is already omnipresent in modern western life (e.g., to access information, facilitate social interactions (social media), and operate security systems). While AI is beginning to be leveraged in clinical settings (e.g., medical imaging, genetic testing) [8], we are still far from routine adoption of AI in healthcare, as the stakes (and potential risks) are much greater than those of the AI that facilitates our modern-day conveniences [9].

AI in Healthcare

AI is currently being used to facilitate early disease detection, enable better understanding of disease progression, optimize medication/treatment dosages, and uncover novel treatments [8, 10–13, 14••, 15]. A major strength of AI is rapid pattern analysis of large datasets. Areas of medicine most successful in leveraging pattern recognition include ophthalmology, cancer detection, and radiology, where AI algorithms can perform as well or better than experienced clinicians in evaluating images for abnormalities or subtleties undetectable to the human eye (e.g., gender from the retina) [16–19]. While it is unlikely that intelligent machines would ever completely replace clinicians, intelligent systems are increasingly being used to support clinical decision-making [8, 14••, 20]. While human learning is limited by capacity to learn, access to knowledge sources, and lived experience, AI-powered machines can rapidly synthesize information from an unlimited amount of medical information sources. To optimize the potential of AI, very large datasets are ideal (e.g., electronic health records; EHRs) that can be analyzed computationally, revealing trends and associations regarding human behaviors and patterns [21] that are often hard for humans to extract.

AI in Mental Healthcare

While AI technology is becoming more prevalent in medicine for physical health applications, the discipline of mental health has been slower to adopt AI [8, 22]. Mental health practitioners are more hands-on and patient-centered in their clinical practice than most non-psychiatric practitioners, relying more on "softer" skills, including forming relationships with patients and directly observing patient behaviors and emotions [23]. Mental health clinical data is often in the form of subjective and qualitative patient statements and written notes. However, mental health practice still has much to benefit from AI technology [24–28]. AI has great potential to re-define our diagnosis and understanding of mental illnesses [29•]. An individual's unique bio-psycho-social profile is best suited to fully explain his/her holistic mental health [30]; however, we have a relatively narrow understanding of the interactions across these biological, psychological, and social systems. There is considerable heterogeneity in the pathophysiology of mental illness and identification of biomarkers may allow for more objective, improved definitions of these illnesses. Leveraging AI techniques offers the ability to develop better prediagnosis screening tools and formulate risk models to determine an individual's predisposition for, or risk of developing, mental illness [27]. To implement personalized mental healthcare as a long-term goal, we need to harness computational approaches best suited to big data.

Machine Learning for Big Data Analysis

Machine learning (ML) is an AI approach that involves various methods of enabling an algorithm to learn [27, 29•, 31–35]. The most common styles of "learning" used for healthcare purposes include supervised, unsupervised, and deep learning [13, 36–38]. There are other ML methods like semi-supervised learning (blend of supervised and unsupervised) [39, 40] and reinforcement learning where the algorithm acts as an agent in an interactive environment that learns by trial and error using rewards from its own actions and experiences [41].

Supervised Machine Learning (SML) Here data are pre-labeled (e.g., diagnosis of major depressive disorder (MDD) vs. no depression) and the algorithm learns to associate input features derived from a variety of data streams (e.g., sociodemographic, biological and clinical measures) to best predict the labels [36, 42]. Labels can be either categorical (MDD or not) or continuous (along a spectrum of severity). The machine experiences SML because the labels act as a "teacher" (i.e., telling the algorithm how to label the data) for the algorithm the "learner" (i.e., learns to associate features with a specific label). After learning from a large amount of labeled *training* data, the algorithm is tested on unlabeled *test* data to determine if it can correctly classify the target variable—e.g., MDD. If the model performance (accuracy or other metric) drops with the test data, the model is considered overfit (recognizing spurious patterns) and cannot be

generalized to external, independent samples. There are algorithms that lend themselves well to SML; some are borrowed directly from traditional statistics like logistic and linear regression, while others are unique to SML like support vector machines (SVM) [43].

Unsupervised Machine Learning (UML) Here algorithms are not provided with labels; thus, the algorithm recognizes similarities between input features and discovers the underlying structure of the data, but is not able to associate features with a known label [37]. UML uses clustering techniques (e.g., k-means, hierarchical, principal component analysis) to sort and separate data into groups or patterns or identify the most salient features of a dataset [44]. The data output must be interpreted by subject-matter experts to determine its usefulness. The lack of labels makes UML more challenging, but able to reveal the underlying structure in a dataset with less a priori bias. For example, neuroimaging biomarkers provide large feature datasets that may hold information regarding unknown subtypes of psychiatric illnesses like schizophrenia. UML may help to identify clusters of biomarkers that characterize these subtypes, thus informing prognosis and best treatment practices.

Deep Learning (DL) DL algorithms learn directly from raw data without human guidance, providing the benefit of discovering latent relationships [45]. DL handles complex, raw data by employing artificial neural networks (ANNs; computer programs that resemble the way a human brain thinks) that process data through multiple "hidden" layers [13, 38, 46]. Given this resemblance to human thinking, DL has been described as less robotic than traditional ML. To be considered "deep," a ANN must have more than one hidden layer [38]. These layers are made up of nodes that combine data input with a set of coefficients (weights) that amplify or dampen that input in terms of its effect on the output. DL is ideal for discovering intricate structures in high-dimensional data like those contained in clinician notes in EHRs [45], or clinical and non-clinical data provided by patients [47, 48]. An important caution in DL is that the hidden layers within ANNs can render the output harder to interpret (black box phenomenon where it is unclear how an algorithm arrived at an output) [49].

Natural Language Processing (NLP) NLP is a subfield of AI that involves using the aforementioned algorithmic methods; however, it specifically refers to how computers process and analyze human language in the form of unstructured text and involves language translation, semantic understanding, and information extraction [50]. Mental health practice will rely heavily on NLP, prior to being able to perform other AI techniques, due to considerable raw input data in the form of text (e.g., clinical notes; other written language) and conversation (e.g., counseling sessions) [48, 51]. The ability of a computer algorithm to automatically understand meanings of underlying words, despite the generativity of human language, is a huge advancement in technology and essential for mental healthcare applications [52].

Analytic Approaches of Traditional Statistical Programming Versus ML

ML methods identify patterns of information in data that are useful to predict outcomes at the individual patient level and do not distinguish samples and populations. The descriptive aspect of statistics is similar to ML, but the inferential aspect, which is the core of statistics, is different, as it uses only samples to make inference about the population from which the sample is drawn [27, 29, 31–35]. Modern ML approaches offer benefits over traditional statistical approaches because they can detect complex (non-linear), high-dimensional interactions that may inform predictions [53–56]. However, the lines between traditional statistics and ML can be blurry due to the overlapping use of analytic approaches [57]. Table 1 summarizes key comparisons between the primary goals of the two approaches. These are only generalizations, as there can be overlap, and should be interpreted as such.

Methods: Study Selection and Performance Measures

Study Selection

To focus this review on recently published literature, we included only studies published 2015–2019, corresponding to the upsurge in AI publications pertaining to mental health (Fig. 1). This is not a systematic review and does not include an exhaustive list of all published studies meeting these broad criteria. We used PubMed and Google Scholar to locate studies that conducted original clinical research in an area relevant to AI and mental health. We did not include studies that described a potential application of AI or development of an algorithm or system that had not yet been tested in a real-world application. We also did not include studies of neurocognitive disorders (e.g., dementia, mild cognitive impairment), despite their relationship to mental health, because there are a considerable number of AI and neurocognition studies that warrant their own review. This review includes a total of 28 original research studies of AI and mental health.

Description of Studies and Performance Metrics Used

We organized Table 2 (details of the 28 studies) based on the nature of the predictive variables used as input for the AI algorithms. The columns summarize the primary study goal, location and population, sample size, mean age, predictors

Table 1 Key comparisons between machine learning and traditional statistics in healthcare research

	Machine learning	Traditional statistics
Year conceptualized	1959	Seventeenth century
Primary goal	Make the best prediction and/or recognize patterns within data (either samples of or an entire study population of interest)	Describe data (samples only) and estimate parameters of an analytic model specified for a population of interest (aka statistical inference)
Knowledge of potential relationships between variables	Not required	Not required for description of data, but required for statistical inference
Hypotheses	More often hypothesis--generating	More often hypothesis-driven
Analysis approach	Often learns from data and models can be difficult to interpret due to extensive use of latent variables (DL and UML black box phenomenon)	Explicitly specified analytic models for statistical inference and easy to interpret
Data size	Very large and can be the size of an entire population of interest	Small to moderate and samples of a population of interest only for statistical inference
Number of features	Large and unspecified	Small and explicitly specified for statistical inference
Rigor	Minimal model assumptions	Strict model assumptions for statistical inference
Interpretability	Limited to data at hand (either example or population) and results	Inference of relationships for the entire population of interest
Methods for assessing performance	Often empirically using cross-validation, ROC AUC, % accuracy, sensitivity, and specificity	Statistical and practical significance (e.g., *p* values, effect sizes)

AUC area under the curve, *DL* deep learning, *ROC* receiver operating characteristic, *UML* unsupervised machine learning

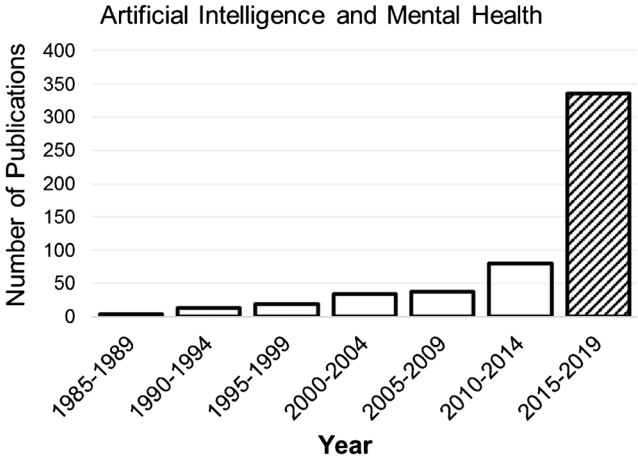

Fig. 1 Frequency of publications by year in PubMed using search terms "artificial intelligence and mental health"

that served as input data, type of AI algorithm and validation, best performing results, and a brief conclusion for each study. Across studies, the most commonly reported performance metrics were as follows:

1. Receiver operating characteristic (ROC) curve. The area under the ROC curve (called AUC), plotted as the true-positive rate (TPR) on the *y*-axis and false-positive rate (FPR) on the *x*-axis [86–89]. The higher the AUC, the better the algorithm is at classifying (e.g., disease vs. no disease); thus, an AUC = 1 indicates perfect ability to distinguish between classes, an AUC = 0.5 means no ability to distinguish between classes (complete overlap), and an AUC = 0 indicates the worst result—all incorrect assignments.
2. Percent (%) accuracy. Percent accuracy is the proportion of correct predictions, determined by dividing the number of correct predictions (true positives + true negatives; TPs + TNs) by all observations (TPs + TNs + false positives and false negatives (FPs + FNs)) [88]. This metric is inadequate, however, when there is uneven class distribution (i.e., significant disparity between the sample sizes for each label).
3. Sensitivity and specificity. Sensitivity is synonymous with the TPR and "recall" (R) and measures the proportion of TPs that are correctly identified (TPs/(TPs + FNs)) [90]. Specificity is synonymous with TNR and measures the proportion of TNs that are correctly identified (TNs/(TNs + FPs)). Sensitivity and specificity are often inversely proportional; as sensitivity increases, specificity decreases and vice versa.
4. Precision (also called positive predictive value; PPV) and F1 scores. Precision is the proportion of *positive* identifications (e.g., presence of MDD) that are correctly classified by the algorithm (TPs/(TPs + FPs)) [86, 91]. For example, precision = 0.5 means that the algorithm correctly predicted MDD 50% of the time. An F1 score is a measure of an algorithm's accuracy that conveys the balance between precision and recall, calculated as 2 * ((precision * recall)/(precision + recall)) [92]. The best value of an F1 score is 1 and the worst is 0. F1 scores can be more useful than accuracy in studies with uneven class distributions.

Table 2 Summary of Studies of AI and Mental Health

Authors	Primary goal	Study setting	Subjects	Sample size (*n*)	Age range
Clinical assessments					
A. Electronic health record (EHR) data					
Arun et al., 2018 [58]	Predict depression (Euro-depression inventory) from hospital EHR data	South India Research Unit, CSI Holdworth Memorial Hospital, Mysuru	Persons born between 1934 and 1966, from the MYNAH database	270 patient records	27 to 67 years
Choi et al., 2018 [59]	Predict probability of suicide death using health insurance records	S. Korea Random sample from health insurance registry of all S. Korean residents	Subjects in the National Health Insurance Service (NHIS)–Cohort Database from 2004 to 2013; Suicide deaths based on ICD-10 codes	819,951 (573,965 training; 1782 suicide deaths) (245,986 testing; 764 suicide deaths)	14+ years. of age
Fernandes et al., 2018 [60]	Detect suicide ideation and attempts using NLP from EHRs	South London (Lambeth, Southwark, Lewisham, and Croydon)	The Clinical Record Interactive Search (CRIS) system from the South London and Maudsley (SLaM) NHS Trust	500 events and correspondence documents	Not reported
Jackson et al., 2016 [61]	Identify symptoms of SMI from clinical EHR text using NLP			23,128 discharge summaries from 7962 patients with SMI; 13,496 discharge summaries from 7575 non-SMI patients	Not reported
Kessler et al., 2017 [62]	Identify veterans at high suicide risk from EHRs	Harvard medical school Data from: US Veterans Health Administration (VHA)	National Death Index (NDI; CDC and Department of HHS, 2015) as having died by suicide in fiscal years 2009–2011	6360 cases 2,108,496 controls (1% probability sample)	Not reported
Sau and Bhakta 2017 [63]	Predict depression from sociodemographic variables and clinical data	Kolkata, India Bagbazar Urban Health and Training Centre	Older adults (43% F) living in a slum with or without depression based on GDS score	105	66.6 ± 5.6 years
B. Mood rating scales					
Chekroud et al., 2016 [64]	Predict whether patients with depression will achieve symptomatic remission after a 12-week course of citalopram	Training data from Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial Testing data from Combining Medications to Enhance Depression Outcomes (CO-MED) trial	Adults with MDD per DSM-IV criteria. Remission based on QIDS-SR follow-up score	Training $n = 1949$ Test $n = 151$	18 to 75 years
Chekroud et al., 2017 [65]	Determine efficacy of antidepressant treatments on empirically defined groups of symptoms and examine replicability of these groups		Adults with MDD based on DSM-IV diagnosis. Training: 63% F Testing: 66% F	Training $n = 4039$ Testing $n = 640$	Training 41.2 ± 13.3 years Testing 42.7 ± 12.2 years
Zilcha-Mano et al., 2018 [66]	Predict who will respond better to placebo or medication in	15 clinical sites in the USA (8-week placebo-controlled RCT trial of citalopram)	Community-dwelling older adults (75+ years) with unipolar depression, diagnosed by HAM-D	174	79.6 ± 4.4 years

Table 2 (continued)

	antidepressant (citalopram) trials		Trajectory based on weekly HAM-D scores (58% F)		
Research assessments					
C. Brain imaging					
Drysdale et al., 2017 [67]	Diagnose subtypes of depression with biomarkers from fMRI	5 academic sites in the USA and Canada	Adults with MDD based on DSM-IV (59% F) HC (58% F)	1188 Training (711; $n = 333$ with depression; $n = 378$ HC Test (477; $n = 125$ depression; $n = 352$ HC)	Training mean = 40.6 years depression Mean = 38.0 years HC
Kalmady et al., 2019 [68]	Classify SZ using fMRI data	National Institute of Mental Health and Neurosciences (NIMHANS, India)	SZ who are antipsychotic drug-naïve and met DSM-IV criteria Age- and sex-matched HC	81 SZ 93 HC	Not reported
Dwyer et al., 2018 [69]	Identify neuro-anatomical subtypes of chronic SZ; determine if subtypes enhance computer-aided discrimination of SZ from HC	Publicly available US data repository of the Mind Research Network and the University of New Mexico	Adults with SZ based on DSM-IV (20% F) HC (31% F)	$n = 71$ schizophrenia $n = 74$ HC 316 test set	38.1 ± 14 years schizophrenia 35.8 ± 11.6 HC
Nenadic et al., 2017 [70]	Detect accelerated brain aging in SZ compared to BD or HC using BrainAGE scores	Jena University, Jena, Thuringia, Germany	Adults with SZ or BD type I based on DSM-IV criteria and HCs (42% F)	137 (45 SZ, 22 BD, 70 HC)	Sz mean 33.7 ± 10.5, (21.4–64.9); HC mean 33.8 ± 9.4, (21.7–57.8); BP mean 37.7 ± 10.7, (23.8–57.7)
Patel et al., 2015 [71]	Predict late-life MDD diagnosis from multimodal imaging and network-based features	Pittsburgh, PA, USA MRI laboratory	Late-life MDD based on DSM-IV criteria age-matched HC	68 Depression $n = 33$ HC $n = 35$	Not reported
Cai et al., 2018 [72]	Classify depression vs. HC using EEG features from the prefrontal cortex	China Laboratory setting (quiet room)	Existing psycho-physiological database of 250 individuals	213 (92 with clinically diagnosed depression 121 HC)	Not reported
Erguzel et al., 2016 [73]	Classify unipolar MDD and BD using EEG data	Istanbul, Turkey	Out-patients with MDD or BD, based on DSM-IV criteria. Medication free for ≥ 48 h	89	Not reported
D. Novel monitoring system					
Bain et al., 2017 [74]	RCT of medication adherence in subjects with SZ using a novel AI platform AiCure	10 US sites Medication adherence in a 24-week clinical trial of drug ABT-126 (ClinicalTrials.gov [NCT01655680])	Stable adult out-patients with SZ who do not smoke (45% F)	75 (53 monitored with AI platform 22 monitored with directly observed therapy)	45.9 ± 10.9 years
Kacem et al., 2018 [75]	Predict depression severity from automated assessments of psychomotor retardation using video data	Not reported Recruited from a clinical trial of depression treatment	Adults with MDD based on DSM-IV criteria Depression severity based on HAM-D scores (60% F)	126 sessions from 49 participants: (56 moderate-severely depressed, 35 mildly depressed, and 35 remitted)	Not reported

Table 2 (continued)

Chattopadhyay 2017 [76]	Mathematically model how psychiatrists clinically perceive depression symptoms and diagnose depression states	India Hospital	Adults with depression, based on DSM-IV criteria. Depression severity rated by clinicians HC—not described	302 depression 50 HC	19 to 50 years
Wahle et al., 2016 [77]	Identify subjects with clinically meaningful depression from smartphone data	Zürich, Switzerland Community smartphone usage over 4 weeks	Clinically depressed adults from Switzerland and Germany Change in depression based on PHQ-9 scale (64% F)	28 (64% F)	20 to 57 years
E. Social media					
Cook et al., 2016 [78]	Predict suicide ideation and heightened psychiatric symptoms from survey data and text messaging data	Madrid, Spain Community text message data over 12 months	Adults (65% F) with recent hospital-based treatment for self-harm who endorsed suicidal ideation (SI) OR did not endorse SI based on text or GHQ-12	1453 $n = 609$ never suicidal $n = 844$ suicidal at some point half of data used for training; half testing	40.5 years 40.0 ± 13.8 years never suicidal 41.6 ± 13.9 years suicidal
Aldarwish and Ahmed 2017 [79]	Identify social network users with depression based on their posts	Saudi Arabia	Posts from Saudi Arabian social network users Training: Depressed post if ≥ 1 DSM-IV MDD symptom mentioned Testing: Depressed subject based on BDI-II scale	Training set = 6773 posts (2073 depressed, 2073 not depressed) Testing set = 30 (15 depressed, 15 not depressed)	Not reported
Deshpande et al., 2017 [80]	Classify Tweets which demonstrate signs of depression and emotional ill-health from those that do not	Twitter API platform	Tweets from all over the world collected at random Categorized based on a curated word list that suggest poor mental health	10,000 Tweets (8000 training; 2000 test)	Not reported
Dos Reis and Culotta 2015 [81]	Detect mood from Twitter data and examine effect of users' physical activity on mental health	Twitter platform	Twitter users who are: Physically active, based on hashtags for activity tracking apps Control users who are not active	$n = 1161$ active $n = 1161$ controls Matched based on gender, location, and online activity	Not reported
Gkotsis et al., 2017 [82]	Classify mental health-related Reddit posts according to theme-based subreddit (topic-specific) groupings	Reddit dataset from (https://www.reddit.com/dev/api) from 1/1/2006 to 8/31/2015	Reddit users	80/20 training/testing split 348,040 users 458,240 mental health-related posts 476,388 non mental health-related posts	Not reported
Mowery et al., 2016 [83]	Classify whether a Twitter post represents evidence of depression and depression subtype	Depressive symptoms and psycho-social stressors associated with depression (SAD) dataset	User information not reported Tweets classified using linguistic annotation scheme based on DSM-5 and DSM-IV criteria	9300 tweets queried using a subset of the Linguistic Inquiry Word Count	Not reported

Table 2 (continued)

Ricard et al., 2018 [84]	Predict depression from community-generated vs. individual-generated social media content	Dartmouth Hanover, NH Clickworker crowd-sourcing platform	Participants on the Clickworker crowd-sourcing platform MDD by PHQ-8 (69% F)	749 10% (78/749) held out as a test set	26.7 ± 7.29 years
Tung and Lu 2016 [85]	Predict depression tendency from web posts	PTT online discussion forum	Chinese web posts between 2004 and 2011	724 posts selected as training/test data Annotated as T/F for depression tendency	Not reported

Authors	Predictors	AI Method SML UML DL NLP CompV	Validation cv	 In sample test	 Out of sample test	Best algorithm/performance	Primary conclusions
Clinical assessments							
A. Electronic health record (EHR) data							
Arun et al., 2018 [58]	Depression, physical frailty, pulmonary function, BMI, LDL	SML	X			XGBoost: accuracy = 98% SN = 98% SP = 94%	Clinical measures can be used to distinguish whether someone has depression
Choi et al., 2018 [59]	Baseline sociodemographic, ICD-10 coded medical conditions	SML DL Stats: Cox regression	X	X		Cox regression: training AUC = 0.72 Testing: AUC = 0.69	Male sex, older age, lower income, medical aid insurance, and disability were linked with suicide deaths at 10-year follow-up
Fernandes et al., 2018 [60]	Suicide terms from epidemiology literature, documents of patients with past suicide attempts, clinician suggestions	NLP SML				SVM: Suicide ideation SN = 88% precision = 92% Suicide attempts SN = 98% Precision = 83%	NLP approaches can be used to identify and classify suicide ideation and attempts in EHR data
Jackson et al., 2016 [61]	SMI symptoms identified by a team of psychiatrists	NLP SML	X			Extracted data for 46 symptoms with a median F1 = 0.88 precision=90% recall = 85% Extractedsymptoms in 87% patients with SMI and 60% patients with non-SMI diagnosis	NLP approaches can be used to extract psychiatric symptoms from EHR data
Kessler et al., 2017 [62]	VHA service use, sociodemographic variables	SML	X		X	Previous study McCarthy et al. (2015) BART: best sensitivity	A different ML model can predict suicide based on fewer

Table 2 (continued)

		Stats: penalized logistic regression model			11% of suicides occurred among 1% of veterans with highest predicted risk and 28% among 5% with highest predicted risk	predictors than the McCarthy 2015 model
Sau and Bhakta 2017 [63]	Sociodemographic and physical comorbidities	DL	X		ANN: Accuracy = 97% AUC = 0.99	Sociodemographic and comorbid conditions can be used to predict the presence of depression
B. Mood rating scales						
Chekroud et al., 2016 [64]	Sociodemographic variables, psychiatric history, mood ratings	SML	X	X	GBM: Training accuracy = 65% AUC = 0.70 $p < 9 \cdot 8 \times 10^{33}$ SN = 63% remission and SP = 66% non-remitters test accuracy = 60%, $p < 0.04$ Escitalopram + buproprion accuracy = 60%, $p < 0.02$ Venlafaxine + mirtazapine accuracy = 51%, $p = 0.53$	Model based on clinical history, sociodemographics, and mood can predict which patients with MDD will respond and remit after taking citalopram
Chekroud et al., 2017 [65]	Items from the QIDS-SR and HAM-D	UML SML	X	X	3 clusters in QIDS-SR (core emotional, insomnia, and atypical symptoms) identified in training 3 clusters replicated in testing GBM: sleep symptom cluster most predictable ($R^2 = 20\%$; $p < 0.01$) Antidepressants (8 of 9) more effective for core emotional symptoms than for sleep or atypical symptoms	3 patient clusters (based on type of depressive symptoms) had varied responses to different antidepressants
Zilcha-Mano et al., 2018 [66]	Sociodemographic, baseline depression, anxiety, cognition, IADLs	SML	X		Medication superior for those with ≤ 12 years education and longer duration depression (> 3.57 years) ($B = 2.5$, $t(32) = 3.0$, $p = 0.004$). Placebo best for those with > 12 years education; almost outperformed medication ($B = -0.57$, $t(96) = -1.9$, $p = 0.06$)	Patients with less education and longer duration of depression more likely to respond to citalopram (than placebo). Patients with more education more likely to respond to placebo (than citalopram)

Research assessments

C. Brain imaging

Table 2 (continued)

Drysdale et al., 2017 [67]	Connectivity in limbic and frontostriatal networks from fMRI data	UML SML	X	X	UL: 4-cluster solution SVM: training accuracy = 89% SN = 84–91% and SP = 84–93% test: accuracy = 86%	Different patterns of fMRI connectivity may distinguish biotypes of MDD with different clinical features and responsiveness to TMS therapy
Kalmady et al., 2019 [68]	Regional activity and functional connectivity from fMRI data	SML	X		Ensemble model: accuracy = 87% (vs. chance 53%)	fMRI measures can distinguish between SZ and HC
Dwyer et al., 2018 [69]	Brain volume measures based on structural MRI data	UML SML	X	X	UL: 2 subgroups SVM: training subgroup improved accuracy 68%–73% (subgroup 1) and 79% (subgroup 2) testing: accuracy decreased: 64%–71% (subgroup 1) and 67% (subgroup 2)	Two neuro-anatomical subtypes of SZ have distinct clinical characteristics, cognitive and symptom courses
Nenadic et al., 2017 [70]	Structural MRI data	SML: Stats: ANOVA			RVR: no accuracy reported significant effect of group on BrainAGE score (ANOVA, $p = 0.009$) SZ had higher mean BrainAGE score than both BD and HC	Using a brain aging algorithm derived from structural MRI data, different diagnostic groups can be compared
Patel et al., 2015 [71]	Multimodal MRI data (functional connectivity, atrophy, integrity, lesions)	SML	X		ADTree: diagnosis accuracy = 87% treatment response accuracy = 89%	MRI data can distinguish late-life depression patients from HCs
Cai et al., 2018 [72]	EEG data while at rest and with sound stimulation	SML UML	X		KNN: average accuracy = 77%	EEG patterns can distinguish between persons with depression and HCs
Erguzel et al., 2016 [73]	EEG data over 12-h period	SML	X		ANN: AUC = 0.76 no feature selection AUC = 0.91 with feature selection	EEG patterns may distinguish between MDD and BD patients
D. Novel monitoring system						
Bain et al., 2017 [74]	Medication adherence monitored by modified directly observed therapy (mDOT)	CompV			90% adherence for subjects monitored with AI platform; 72% for subjects monitored by mDOT	A novel AI platform has better medication adherence than directly observed therapy in persons with SZ
Kacem et al., 2018 [75]	Measurement of face and head motion based on video recordings	SML	X		SVM: accuracy of facial movement = 66% head movement = 61% Combined = 71% Highest accuracy for severe vs. mild depression 84%	Facial (but not head) movements may be used to distinguish severity levels of depression
		DL	X			

Table 2 (continued)

Chattopadhyay 2017 [76]	Psychiatrists' ratings of individual symptoms					Fuzzy neural hybrid model: accuracy = 96%	The link between clinicians' assessments of symptoms and overall depression severity can be modeled by AI
Wahle et al., 2016 [77]	Smartphone usage, accelerometer, Wi-Fi, and GPS data (movement, activity)	SML	X			RF accuracy = 62%	Smartphone sensor data can distinguish between those with and without depression at follow-up
E. Social media							
Cook et al., 2016 [78]	Survey (sleep, depressive symptoms, medications), and unstructured data: text response to "how are you feeling today?"	NLP SML			X	Multivariate logistic model: suicide ideation (structured better) PPV = 0.73, SN = 0.76, SP = 0.62 Heightened psychiatric symptoms (structured better) PPV = 0.79, SN = 0.79, SP = 0.85	NLP-based models of unstructured texts have high predictive value for SI, and may require less time and effort from subjects
Aldarwish and Ahmed 2017 [79]	Social network posts from LiveJournal, Twitter, and Facebook	NLP SML				NB: Accuracy = 63% Precision = 100% Recall = 58%	Social media posts could be used to identify which users are depressed
Deshpande et al., 2017 [80]	Unstructured text (Tweet)	NLP SML		X		NB: Precision = 0.836 Accuracy = 83%	Text-based emotion can detect depression from Twitter data
Dos Reis and Culotta 2015 [81]	2367 unstructured text (Tweets) that were hand classified as expressing either anxiety, depression, anger, or none	SML Stats: Wilcoxon singed rank	X			Logistic regression classifier: hostility AUC = 0.901 Dejection AUC = 0.870 Anxiety AUC = 0.850 physically active users had 2.7% fewer anxious tweets and 3.9% fewer dejected tweets than a matched user	Social media posts can be used to infer negative mood states. Physically active social media users post fewer Tweets reflecting negative mood states
Gkotsis et al., 2017 [82]	Identified subreddits related to mental health using keywords	Semi-SML DL		X		CNN: accuracy = 91.08% distinguishing mental health posts Precision = 0.72 Recall = 0.71 for which theme a post belonged to	Can distinguish mental health-related Reddit posts from unrelated posts as well as the mental health theme they relate to; identified 11 mental health themes
Mowery et al., 2016 [83]	Unstructured text (Tweet)	NLP SML	X			SVM: F1 score = 52 for a tweet with evidence of depression	Text analysis of tweets can be used to identify depressive symptoms and subtype
		NLP	X	X		Training: not reported	

Table 2 (continued)

Ricard et al., 2018 [84]	Unstructured text data (Instagram posts and comment), demographics, other survey data	SML	Testing: Elastic-net RLR model: community-generated AUC = 0.71, $p < 0.03$ Combination AUC = 0.72, $p < 0.02$ User-generated AUC = 0.63, $p = 0.11$	Instagram posts (both user-generated and community-generated content) can distinguish people with depression
Tung and Lu 2016 [85]	Unstructured text data (posts)	NLP	EDDTW highest recall = 0.67 and *F* measure = 0.62 DSM precision = 0.666	NLP of web posts can identify depressive tendencies

ADTree alternating decision tree, *ANN* artificial neural network, *BAO* Beck Anxiety Inventory, *BDI* Beck Depression Inventory, *cTAKES* clinical text analysis knowledge extract system, *CompV* computer vision, *DL* deep learning, *EDDTW* event-driven depression tendency warning, *GAD-7* generalized anxiety disorder, *GHQ-12* General Health Questionnaire, *GMM* Gaussian mixture models, *HAM-D* Hamilton Rating Scale for Depression, *HC* healthy control, *HHS* health and human services, *JSON* JavaScript Object Notation, *LES* life event scale, *LDA* linear discriminant analysis, *MDD* major depressive disorder, *MMSE* Mini-Mental State Examination, *NLP* natural language processing, *PANSS* Positive and Negative Syndrome Scale, *PHQ-9* Patient Health Questionnaire, *PPV* positive predictive value, *PSQI* Pittsburgh Sleep Quality Index, *QIDS-SR* Quick Inventory of Depressive Symptomatology, *SL* supervised learning, *SMI* severe mental illness, *SN* sensitivity, *SP* specificity, *SCID-I* Structured Clinical Interview for Axis I Disorders, *SVM* support vector machine, *UL* unsupervised learning

Results: Summary of Mental Health Literature

Summary of AI Studies of Mental Health

We categorized Table 2 by the nature of the predictor variables used as input data, including the following: A, electronic health records (EHRs) (6/28) [58–61, 62•, 63]; B, mood rating scales (3/28) [64•, 65•, 66]; C, brain imaging data (7/28) [67•, 68, 69•, 70–73]; D, novel monitoring systems (e.g., smartphone, video) (4/28) [74–77]; and E, social media platforms (e.g., Twitter) (8/28) [78–81, 83–85]. Depression (or mood) was the most common mental illness investigated (18/28) [58, 63–67, 71–73, 75–77, 79–81, 83–85]. We also found examples of AI applied to schizophrenia and other psychiatric illnesses (6/28) [61, 68–70, 73, 74], suicidal ideation/attempts (4/28) [59, 60, 62•, 78], and general mental health (1/28) [82]. Participants included in these studies were either healthy controls or were diagnosed with a specified mental illness. Sample sizes ranged from small ($n = 28$) [77] to large ($n = 819{,}951$) [59]. There was no age reported for 14/28 studies likely due to the nature of the data (e.g., social media platform or other anonymous database). For the remainder, ages ranged from 14+ years [59] to a mean age of 79.6 (SD 4.4) years [66].

SML was the most common AI technique (23/28), and a proportion of studies (8/28) also used NLP prior to applying ML. Cross-validation techniques were most common (19/28), but several studies also tested the algorithm on a held-out subsample not used for training (4/28), or in an external validation sample (6/28). There was considerable heterogeneity in the nature of the results reported across studies. Accuracies ranged from the low 60s (62% from smartphone data [77] and 63% from social media posts [79]) to high 90s (98% from clinical measures of physical function, body mass index, cholesterol, etc. [58] and 97% from sociodemographic variables and physical comorbidities [63]) for prediction of depression. ML methods were also able to predict treatment responses to commonly prescribed antidepressants like citalopram (65% accuracy) [64], or identify features like education that were related to placebo versus medication responses [66].

NLP techniques identified symptoms of severe mental illness from EHR data (precision = 90%; recall = 85%) [61]. Brain MRI features identified neuroanatomical subtypes of schizophrenia with 63–71% accuracy [69•], and fMRI features classified schizophrenia (vs. controls) with 87% accuracy [68]. An AI platform resulted in more successful medication adherence for patients with schizophrenia (90%) than modified directly observed therapy (72%) [74]. Health insurance records (AUC = 0.69) [59], survey and text message data (sensitivity = 0.76; specificity = 0.62) [78], and EHRs (suicidal ideation; sensitivity = 88%; precision = 92% and suicide attempts; sensitivity = 98%; precision = 83%) [60] all enabled prediction of suicidal ideation and attempts.

Limitations of AI and Mental Health Studies

These studies have limitations pertaining to clinical validation and readiness for implementation in clinical decision-making and patient care. As recognized for any AI application, the size and quality of the data limit algorithm performance [13]. For small sample sizes, overfitting of the ML algorithms is highly likely [28]. Testing the ML models only within the same sample and not out-of-sample limits the generalizability of the results. The predictive ability of these studies is restricted to the features (e.g., clinical data, demographics, biomarkers) used as input for the ML models. As no one study is exhaustive in this manner, the clinical efficacy of the particular features used to derive these models must be considered. It is also possible that the outputs of these algorithms are only valid under certain situations or for a certain group of people. These studies were not always explicitly clear regarding the significance or practical meaning of resulting performance metrics. For example, performance accuracy should be compared to clinical diagnostic accuracy (as opposed to simply relating these values to chance) in order to interpret clinical value [89].

The use of binary classifiers is more common in ML than regression models (i.e., continuous scores) due to being easier to train; however, a consequence of this approach is overlooking the severity of a condition [32]. Future studies should seek to model severity of mental illnesses along a continuum. While these studies focused on features that are considered risk factors for mental illnesses, subsequent research should also consider investigating protective factors like wisdom that can improve an individual's mental health [93, 94]. Finally, a challenge in studies seeking to model rare events (e.g., suicide) or illness is that of highly imbalanced datasets (i.e., the event rarely occurs or a relatively small portion of the population develops the illness). In these instances, classifiers tend to predict outcomes as the majority class (e.g., miss rare events like suicide ideation) [95]. Techniques employed in these studies to overcome this challenge included (i) under-sampling (reducing number of samples in the majority) [62•], (ii) over-sampling (matching the ratio of major and minor groups by duplicating samples for the minor group) [59], and (iii) ensemble learning methods (combining several models to reduce variance and improve predictions) [68, 77]; however, few studies (4/28) reported using these techniques.

Discussion: Future Research Directions and Recommendations

The World Health Organization defines health as, "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" [96]. If we leverage today's available technologies, we can obtain continuous, long-term monitoring of the unique bio-psycho-social profiles of individuals [26] that impact their mental health. The resulting amount of complex, multimodal data is too much for a human to process in a meaningful way, but AI is well suited to this task. As AI techniques continue to be refined and improved, it may be possible to define mental illnesses more objectively than the current DSM-5 classification schema [97], identify mental illnesses at an earlier or prodromal stage when interventions may be more effective, and tailor prescribed treatments based on the unique characteristics of an individual.

Areas Needing Additional Research for AI and Mental Health

In order to discover new relationships between mental illnesses and latent variables, very large, high-quality datasets are needed. Obtaining such deeply phenotyped large datasets poses a challenge for mental health research and should be a collaborative priority (e.g., robust platforms for data sharing among institutes). DL methods will be increasingly necessary (over SML methods) to handle these complex data, and the next challenge will be in ensuring that these models are clinically interpretable rather than a "black box" [13, 49, 98]. Transfer learning, where an algorithm created for one purpose is adapted for a different purpose, will help to strengthen ML models and improve their performance [99]. Transfer learning is already being applied to fields that rely heavily on image analysis like pathology, radiology, and dermatology, including commercial efforts to integrate these algorithms in clinical settings [100, 101]. Flexible algorithms will likely be a greater challenge for mental health due to the heterogeneity in salient input data. Additionally, AI models should have a life-long learning framework to prevent "catastrophic forgetting" [102]. Collaborative efforts between data scientists and clinicians to develop robust algorithms will likely yield the best results.

AI algorithms will be developed from emerging data sources, and these data may not be fully representative of constructs of interest or populations. For example, social media data (e.g., "depressive" posts) may not be representative of the construct of interest (depression). Posts containing words indicative of depression could suggest a transient state of depressive mood rather than a diagnosis of depression. Social media posters also may exaggerate symptoms in online posts or their comments could simply be contextual. Thus, the data could be misconstrued due to the limited contextual information [103]. The clinical usefulness of these platforms of rich information requires more careful consideration, and studies using social media need to be held to higher methodological

standards. Finally, the use of AI to derive insights from data may help to facilitate diagnosis, prognosis, and treatment; however, it is important to consider the practicality of these insights and whether they can be translated and implemented in the clinic [89].

How AI Can Benefit Current Healthcare for Individuals with Mental Illnesses

Physician time is progressively limited as mental healthcare needs grow and clinicians are burdened with increased documentation requirements and inefficient technology. These problems are particularly cumbersome for mental health practitioners who must rely on their uniquely human skills in order to foster therapeutic rapport with their patients and design personalized treatments. Use of AI technology offers many benefits in addition to improving detection and diagnosis of mental illnesses. AI algorithms can be harnessed to comprehensively draw meaning from large and varied data sources, enable better understanding of the population-level prevalence of mental illnesses, uncover biological mechanisms or risk/protective factors, offer technology to monitor treatment progress and/or medication adherence, deliver remote therapeutic sessions or provide intelligent self-assessments to determine severity of mental illness, and perhaps most importantly enable mental health practitioners to focus on the human aspects of medicine that can only be achieved through the clinician–patient relationship [5, 20].

Ethical Considerations for AI in Mental Healthcare Practice

To deploy AI responsibly, it is critical that algorithms used to predict or diagnose mental health illnesses be accurate and not lead to increased risk to patients. Moreover, those involved in making decisions about the selection, testing, implementation, and evaluation of AI technologies must be aware of ethical challenges, including biased data (e.g., subjective and expressive nature of clinical text data; linking of mental illnesses to certain ethnicities) [104]. Accepted ethical principles used to guide biomedical research including autonomy, beneficence, and justice must be prioritized and in some cases augmented [105]. It is critical that data and technology literacy gaps be addressed for both patients and clinicians. Moreover, to our knowledge there are no established standards to guide the use of AI and other emerging technologies in healthcare settings [106]. Computational scientists may train AI using datasets that lack sufficient data to make meaningful assessments or predictions [107]. Clinicians may not know how to manage the depth of granular data nor be confident with a decision produced by AI [108]. Institutional Review Boards have limited knowledge of emerging technologies, which makes risk assessment inconsistent [106]. For example, there are efforts to link smartphone keystrokes and voice patterns to mood disorders, and yet the public may not be aware such linkages are possible [109]. Public communication about these algorithms must be useful, contextual, and confer that tools supplement, but do not replace, medical practice. Clearly, there is a need to integrate ethics into the development of AI via research and education and resources will need to be appropriated for this purpose.

Concluding Remarks

AI is increasingly a part of digital medicine and will contribute to mental health research and practice. A diverse community of experts vested in mental health research and care, including scientists, clinicians, regulators, and patients must communicate and collaborate to realize the full potential of AI [110]. As elegantly suggested by De Choudhury et al., a critical element is combining human intelligence with AI to (1) ensure construct validity, (2) appreciate unobserved factors not accounted for in data, (3) assess the impact of data biases, and (4) proactively identify and mitigate potential AI mistakes [111]. The future of AI in mental healthcare is promising. As researchers and practitioners vested in improving mental healthcare, we must take an active role in informing the introduction of AI into clinical care by lending our clinical expertise and collaborating with data and computational scientists, as well as other experts, to help transform mental health practice and improve care for patients.

Funding Information This study was supported, in part, by the National Institute of Mental Health T32 Geriatric Mental Health Program (grant MH019934 to DVJ [PI]), the IBM Research AI through the AI Horizons Network IBM-UCSD AI for Healthy Living (AIHL) Center, by the Stein Institute for Research on Aging at the University of California San Diego, and by the National Institutes of Health, Grant UL1TR001442 of CTSA funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Compliance with Ethical Standards

Conflict of Interest Sarah Graham, Xin Tu, and Ho-Cheol Kim each declare no potential conflicts of interest.

Colin Depp and Dilip V. Jeste are Co-Directors of UCSD-IBM Center on Artificial Intelligence for Healthy Living (2018–2022). This is a grant to UCSD from IBM. Drs. Depp and Jeste have no commercial interest in IBM or any other AI-related companies.

Ellen E. Lee has received grants from The National Institute of Mental Health, The National Institutes of Health, and The Stein Institute for Research on Aging.

Camille Nebeker is a co-investigator on a grant supported by IBM and her research on the ethics of emerging technologies is supported by the Robert Wood Johnson Foundation.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as: • Of importance •• Of major importance

1. Pang Z, Yuan H, Zhang Y-T, Packirisamy M. Guest Editorial Health Engineering Driven by the Industry 4.0 for Aging Society. IEEE J Biomed Heal Informatics. 2018;22(6):1709–10. https://doi.org/10.1109/JBHI.2018.2874081.
2. Schwab K. The fourth Industrial Revolution. First. New York, NY: Currency; 2017. p. 192.
3. Simon HA. Artificial intelligence: where has it been, and where is it going? IEEE Trans Knowl Data Eng. 1991;3(2):128–36. https://doi.org/10.1109/69.87993.
4. Metz C, Smith CS. "A.I. can be a boon to medicine that could easily go rogue'. The New York Times. 2019 Mar 25;B5.
5. Kim JW, Jones KL, Angelo ED. How to prepare prospective psychiatrists in the era of artificial intelligence. Acad Psychiatry. 2019;43:1–3. https://doi.org/10.1007/s40596-019-01025-x.
6. John McCarthy. Artificial intelligence, logic and formalizing common sense. In Philosophical logic and artificial intelligence 1989 (pp. 161-190). Springer, Dordrecht.
7. Turing AM. Computing machinery and intelligence. Comput Mach Intell. 1950;49:433–60 Available from: https://linkinghub.elsevier.com/retrieve/pii/B978012386980750023X.
8. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43. https://doi.org/10.1136/svn-2017-000101.
9. Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. Technol Forecast Soc Chang. 2016;105:105–20. https://doi.org/10.1016/j.techfore.2015.12.014.
10. Beam AL, Kohane IS. Translating artificial intelligence into clinical care. JAMA. 2016;316(22):2368–9. https://doi.org/10.1001/jama.2016.17217.
11. Bishnoi L, Narayan Singh S. Artificial intelligence techniques used in medical sciences: a review. Proc 8th Int Conf Conflu 2018. Cloud Comput Data Sci Eng Conflu. 2018;2018:106–13. https://doi.org/10.1109/CONFLUENCE.2018.8442729.
12. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. Npj Digit Med. 2018;1(1):3–6. https://doi.org/10.1038/s41746-017-0012-2.
13. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2017;19(6):1236–46. https://doi.org/10.1093/bib/bbx044.
14.•• •• Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7 **This review provides a current overview of artificial intelligence applications in all areas of medicine.**
15. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. J R Soc Med. 2019;112(1):22–8. https://doi.org/10.1177/0141076818815510.
16. Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer. 2019;111:30–7. https://doi.org/10.1016/j.ejca.2018.12.016.
17. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–10. https://doi.org/10.1038/s41568-018-0016-5.
18. Sengupta PP, Adjeroh DA. Will artificial intelligence replace the human echocardiographer? Circulation. 2018;138(16):1639–42. https://doi.org/10.1161/CIRCULATIONAHA.118.037095.
19. Vidal-Alaball J, Royo Fibla D, Zapata MA, Marin-Gomez FX, Solans FO. Artificial intelligence for the detection of diabetic retinopathy in primary care: protocol for algorithm development. JMIR Res Protoc. 2019;8(2):e12539. https://doi.org/10.2196/12539.
20. Topol E. Deep medicine: how artificial intelligence can make healthcare human again. 1st ed. New York, NY: Basic Books; 2019.
21. Wang Y, Kung LA, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change. 2016;126:3–13. https://doi.org/10.1016/j.techfore.2015.12.019.
22. Miller DD, Facp CM, Brown EW. Artificial intelligence in medical practice: the question to the answer ? Am J Med. 2018;131(2):129–33. https://doi.org/10.1016/j.amjmed.2017.10.035.
23. Gabbard GO, Crisp-Han H. The early career psychiatrist and the psychotherapeutic identity. Acad Psychiatry. 2017;41(1):30–4. https://doi.org/10.1007/s40596-016-0627-7.
24. Janssen RJ, Mourão-Miranda J, Schnack HG. Making individual prognoses in psychiatry using neuroimaging and machine learning. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018;3(9):798–808. https://doi.org/10.1016/j.bpsc.2018.04.004.
25. Luxton DD. Artificial intelligence in psychological practice: current and future applications and implications. Prof Psychol Res Pract. 2014;45(5):332–9. https://doi.org/10.1037/a0034559.
26. Mohr D, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol. 2017;13:23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949.
27. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med. 2019;49:1–23. https://doi.org/10.1017/S0033291719000151.
28. Iniesta R, Stahl D, Mcguf P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med. 2016;46(May):2455–65. https://doi.org/10.1017/S0033291716001367.
29.• • Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018;3(3):223–30. https://doi.org/10.1016/j.bpsc.2017.11.007 **This review aquaints the reader with key terms related to artificial intelligence and psychiatry and gives an overview of the opportunities and challenges in bringing machine intelligence into psychiatric practice**.
30. Jeste DV, Glorioso D, Lee EE, Daly R, Graham S, Liu J, et al. Study of independent living residents of a continuing care senior housing community: sociodemographic and clinical associations of cognitive, physical, and mental health. Am J Geriatr Psychiatry [Internet]. 2019. https://doi.org/10.1016/j.jagp.2019.04.002.
31. Chen M, Hao Y, Hwang K, Wang L, Access LW-I, 2017. Disease prediction by machine learning over big data from healthcare communities. IEEE Access 2017;5:8869–8879. DOI: https://doi.org/10.1109/ACCESS.2017.2694446.
32. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Sci Mag. 2015;349(6245):255–60. https://doi.org/10.1126/science.aaa8415.
33. Nevin L. Advancing the beneficial use of machine learning in health care and medicine: toward a community understanding. PLoS Med. 2018;15(11):4–7. https://doi.org/10.1371/journal.pmed.1002708.
34. Srividya M, Mohanavalli S, Bhalaji N. Behavioral modeling for mental health using machine learning algorithms. J Med Syst. 2018;42:88. https://doi.org/10.1007/s10916-018-0934-5.

35. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149–53. https://doi.org/10.1093/cid/cix731.
36. Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. Nat Methods. 2018;15(1):5–6. https://doi.org/10.1038/nmeth.4551.
37. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6(26094):1–10. https://doi.org/10.1038/srep26094.
38. LeCun Y, Bengio Y, Hinton G. Deep learning. Nat Methods. 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.
39. Ding S, Zhu Z, Zhang X. An overview on semi-supervised support vector machine. Neural Comput & Applic. 2017;28(5):969–78. https://doi.org/10.1007/s00521-015-2113-7.
40. Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform. 2016;64:168–78. https://doi.org/10.1016/j.jbi.2016.10.007.
41. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25(1):14–8. https://doi.org/10.1038/s41591-018-0310-5.
42. Fabris F, de Magalhães JP, Freitas AA. A review of supervised machine learning applied to ageing research. Biogerontology. 2017;18(2):171–88. https://doi.org/10.1007/s10522-017-9683-y.
43. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng. 2007;160:3–24.
44. Dy JG, Brodley CE. Feature selection for unsupervised learning. J Mach Learn Res. 2004;5:845–89 Retrieved from: http://www.jmlr.org/papers/volume5/dy04a/dy04a.pdf.
45. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Heal Informatics 2018;22(5):1589–1604. DOI: https://doi.org/10.1109/JBHI.2017.2767063.
46. Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. Deep learning for healthcare applications based on physiological signals: a review. Comput Methods Prog Biomed. 2018;161(April):1–13. https://doi.org/10.1016/j.cmpb.2018.04.005.
47. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. Trans Assoc Comput Linguist. 2016;4:463–76. https://doi.org/10.1162/tacl_a_00111.
48. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng. 2017;23(05):649–85. https://doi.org/10.1017/S1351324916000383.
49. Samek W, Wiegand T, Müller K-R. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv Prepr arXiv. 2017;1708.08296. Available from: http://arxiv.org/abs/1708.08296
50. Hirschberg J, Manning CD. Advances in natural language processing. Sci Mag. 2015;349(6245):261–6. https://doi.org/10.1126/science.aaa8685.
51. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760–72. https://doi.org/10.1016/j.jbi.2009.08.007.
52. Cambria E, White B. Jumping NLP curves: a review of natural language processing research. IEEE Comput Intell Mag. 2014;9(2):48–57. https://doi.org/10.1109/MCI.2014.2307227.
53. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15(4):233–4. https://doi.org/10.1038/nmeth.4642.
54. Hand DJ. Statistics and data mining: intersecting disciplines. ACM SIGKDD Explor Newsl. 1999;1(1):16–9. https://doi.org/10.1145/846170.846171.
55. Scott EM. The role of statistics in the era of big data: crucial, critical and under-valued. Stat Probab Lett. 2018;136:20–4. https://doi.org/10.1016/j.spl.2018.02.050.
56. Sargent DJ. Comparison of artificial neural networks with other statistical approaches. Cancer. 2002;91(S8):1636–42. https://doi.org/10.1002/1097-0142(20010415)91:8+<1636::AID-CNCR1176>3.0.CO;2-D.
57. Breiman L. Statistical modeling: the two cultures. Stat Sci. 2001;16(3):199–231. https://doi.org/10.1214/ss/1009213726.
58. Arun V, Prajwal V, Krishna M, Arunkumar BV, Padma SK, Shyam V. A boosted machine learning approach for detection of depression. Proc 2018 IEEE Symp Ser Comput Intell SSCI. 2018;2018:41–7. https://doi.org/10.1109/SSCI.2018.8628945.
59. Choi SB, Lee W, Yoon JH, Won JU, Kim DW. Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. J Affect Disord. 2018;231(January):8–14. https://doi.org/10.1016/j.jad.2018.01.019.
60. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci Rep. 2018;8(1):7426. https://doi.org/10.1038/s41598-018-25773-2.
61. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open. 2017;7(1):e012012. https://doi.org/10.1136/bmjopen-2016-012012.
62.• • Kessler RC, Hwang I, Hoffmire CA, Mccarthy JF, Maria V, Rosellini AJ, et al. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psychiatr Res. 2017;26(3):1–14. https://doi.org/10.1002/mpr.1575 **This study from the US Veterans Health Administration (VHA) compared machine learning approaches within and out of sample with traditional statistics to identify veterans at high suicide risk for more targeted care**.
63. Sau A, Bhakta I. Artificial neural network (ANN) model to predict depression among geriatric population at a slum in Kolkata, India. J Clin Diagn Res. 2017;11(5):VC01–4. https://doi.org/10.7860/JCDR/2017/23656.9762.
64.• • Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3(3):243–50. https://doi.org/10.1016/S2215-0366(15)00471-X **This study used machine learning to identify 25 variables from the STAR*D clinical trial that were most predictive of treatment outcome following a 12-week course of the antidepressant citalopram and externally validated their models in an indepdent sample from the CO-MED clinical trial undergoing escitalopram treatment**.
65.• • Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the efficacy and predictability of antidepressant treatments a symptom clustering approach. JAMA Psychiatry. 2017;74(4):370–8. https://doi.org/10.1001/jamapsychiatry.2017.0025 **This study demonstrated that clusters of symptoms are detectable in 2 common depression rating scales (QIDS-SR and HAM-D), and these symptom clusters vary in their responsiveness to different antidepressant treatments**.
66. Zilcha-Mano S, Roose SP, Brown PJ, Rutherford BR. A machine learning approach to identifying placebo responders in late-life

depression trials. Am J Geriatr Psychiatry. 2018;26(6):669–77. https://doi.org/10.1016/j.jagp.2018.01.001.

67.• • Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med. 1878;23(1):28–38. DOI: https://doi.org/10.1038/nm.4246. **This study used unsupervised and supervised machine learning with fMRI data and demonstrated that patients with depression can be subdivided into four neurophysiological subtypes defined by distinct patterns of dysfunctional connectivity in limbic and frontostriatal networks and further that these subtypes predicted which patients responded to repetitive transcranial magnetic stimulation (TMS) therapy**.

68. Kalmady SV, Greiner R, Agrawal R, Shivakumar V, Narayanaswamy JC, Brown MRG, et al. Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. NPJ Schizophr. 2019;5(1):2. https://doi.org/10.1038/s41537-018-0070-8.

69.• • Dwyer DB, Cabral C, Kambeitz-Ilankovic L, Sanfelici R, Kambeitz J, Calhoun V, et al. Brain subtyping enhances the neuroanatomical discrimination of schizophrenia. Schizophr Bull. 2018;44(5):1060–9. https://doi.org/10.1093/schbul/sby008 **This study used both unsupervised and supervised machine learning with structural MRI data and suggested that sMRI-based subtyping enhances neuroanatomical discrimination of schizophrenia by identifying generalizable brain patterns that align with a clinical staging model of the disorder**.

70. Nenadić I, Dietzek M, Langbein K, Sauer H, Gaser C. BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. Psychiatry Res Neuroimaging. 2017;266(March):86–9. https://doi.org/10.1016/j.pscychresns.2017.05.006.

71. Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF, Aizenstein HJ. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int J Geriatr Psychiatry. 2015;30(10):1056–67. https://doi.org/10.1002/gps.4262.

72. Cai H, Han J, Chen Y, Sha X, Wang Z, Hu B, et al. A pervasive approach to EEG-based depression detection. Complexity. 2018;2018:1–13. https://doi.org/10.1155/2018/5238028.

73. Erguzel TT, Sayar GH, Tarhan N. Artificial intelligence approach to classify unipolar and bipolar depressive disorders. Neural Comput & Applic. 2016;27(6):1607–16. https://doi.org/10.1007/s00521-015-1959-z.

74. Bain EE, Shafner L, Walling DP, Othman AA, Chuang-Stein C, Hinkle J, et al. Use of a novel artificial intelligence platform on mobile devices to assess dosing compliance in a phase 2 clinical trial in subjects with schizophrenia. JMIR mHealth uHealth. 2017;5(2):e18. https://doi.org/10.2196/mhealth.7030.

75. Kacem A, Hammal Z, Daoudi M, Cohn J. Detecting depression severity by interpretable representations of motion dynamics. Proc - 13th IEEE Int Conf Autom Face Gesture Recognition, FG. 2018;2018:739–45. https://doi.org/10.1109/FG.2018.00116.

76. Chattopadhyay S. A fuzzy approach for the diagnosis of depression. Appl Comput Informatics. 2018;13(1):10–8. https://doi.org/10.1016/j.aci.2014.01.001.

77. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. JMIR mHealth uHealth. 2016;4(3):e111. https://doi.org/10.2196/mhealth.5960.

78. Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Comput Math Methods Med. 2016;2016:1–8. https://doi.org/10.1155/2016/8708434.

79. Aldarwish MM, Ahmad HF. Predicting depression levels using social media posts. Proc - 2017 IEEE 13th Int Symp Auton Decentralized Syst ISADS 2017. 2017;277–80. DOI: https://doi.org/10.1109/ISADS.2017.41.

80. Deshpande M, Rao V. Depression detection using emotion artificial intelligence. Proc Int Conf Intell Sustain Syst ICISS. 2017;2017:858–62. https://doi.org/10.1109/ISS1.2017.8389299.

81. Landeiro Dos Reis V, Culotta A. Using matched samples to estimate the effects of exercise on mental health from twitter. Proc Twenty-Ninth AAAI Conf Artif Intell. 2015:182–8 Retrieved from: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewPaper/9960.

82. Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJP, Dobson RJB, et al. Characterisation of mental health conditions in social media using informed deep learning. Sci Rep. 2017;7(1):1–10. https://doi.org/10.1038/srep45141.

83. Mowery D, Park A, Conway M, Bryan C. Towards automatically classifying depressive symptoms from twitter data for population health. Proc Work Comput Model People's Opin Personal Emot Soc Media. 2016:182–91 Available from: https://www.aclweb.org/anthology/W16-4320.

84. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. J Med Internet Res. 2018;20(12):e11817. https://doi.org/10.2196/11817.

85. Tung C, Lu W. Analyzing depression tendency of web posts using an event-driven depression tendency warning model. Artif Intell Med. 2016;66:53–62. https://doi.org/10.1016/j.artmed.2015.10.003.

86. Šimundić A-M. Measures of diagnostic accuracy: basic definitions. Ejifcc. 2009;19(4):203–11.

87. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30(7):1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.

88. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 2005;17(3):299–310. https://doi.org/10.1109/TKDE.2005.50.

89. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis. Radiology. 2018;286(3):800–9. https://doi.org/10.1148/radiol.2017171920.

90. Parikh R, Mathai A, Parikh S, Sekhar C, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008;56(1):45–50.

91. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.

92. Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. Mach Learn Knowl Discov Databases. 2014;8725:225–39. https://doi.org/10.1007/978-3-662-44851-9_15.

93. Lee EE, Depp C, Palmer BW, Glorioso D, Daly R, Liu J, et al. High prevalence and adverse health effects of loneliness in community-dwelling adults across the lifespan: role of wisdom as a protective factor. Int Psychogeriatr. 2018;(May):1–16. https://doi.org/10.1017/S1041610218002120.

94. Jeste DV. Positive psychiatry comes of age. Int Psychogeriatr. 2018;30(12):1735–8. https://doi.org/10.1017/S1041610218002211.

95. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine

learning. J Mach Learn Res. 2017;18(1):559–63 Available from: http://www.jmlr.org/papers/volume18/16-365/16-365.pdf.
96. World Health Organization. Frequently asked questions. 2019. Available from: https://www.who.int/about/who-we-are/frequently-asked-questions
97. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Publication; 2013.
98. Freitas AA. Comprehensible classification models—a position paper. ACM SIGKDD Explor Newsl. 2014;15(1):1–10. https://doi.org/10.1145/2594473.2594475.
99. Torrey L, Shavlik J. Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global. 2009:242–64.
100. Fu G, Levin-schwartz Y, Lin Q, Zhang D, Fu G, Levin-schwartz Y, et al. Machine learning for medical imaging. J Healthc Eng. 2019;2019:10–2. https://doi.org/10.1148/rg.2017160130.
101. Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and future. In: Classification in BioApps. Springer Cham.; p. 323–50.
102. Kemker R, McClure M, Abitino A, Hayes T, Kanan C. Measuring catastrophic forgetting in neural networks. Thirty-second AAAI Conf Artif Intell. 2018:3390–8 Available from: http://arxiv.org/abs/1708.02072.
103. Ruths D, Pfeffer J. Social media for large studies of behavior. Sci Mag. 2014;346(6213):1063–4. https://doi.org/10.1126/science.346.6213.1063.
104. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. 2019;21(2):E167–79. https://doi.org/10.1001/amajethics.2019.167.
105. Raymond N. Safeguards for human studies can't cope with big data. Nature. 2019;568(7752):277. https://doi.org/10.1038/d41586-019-01164-z.
106. Nebeker C, Harlow J, Giacinto RE, Orozco- r, Bloss CS, Weibel N, et al. Ethical and regulatory challenges of research using pervasive sensing and other emerging technologies: IRB perspectives. AJOB Empir Bioeth 2017;8(4):266–276. DOI: https://doi.org/10.1080/23294515.2017.1403980.
107. Sears M. AI Bias and the "people factor" in AI development. 2018 [cited 2019 Feb 26]. Available from: https://www.forbes.com/sites/colehaan/2019/04/30/from-the-bedroom-to-the-boardroom-how-a-sleepwear-company-is-empowering-women/#7717796a2df3
108. Adibuzzaman M, Delaurentis P, Hill J, Benneyworth D. Big data in healthcare—the promises , challenges and opportunities from a research perspective: a case study with a model database. AMIA Annu Symp Proc. 2017;2017:384–92.
109. Huang H, Cao B, Yu PS, Wang C-D, Leow AD. dpMood: exploiting local and periodic typing dynamics for personalized mood prediction. 2018 IEEE Conf Data Min. 2018:157–66. https://doi.org/10.1109/ICDM.2018.00031.
110. Özdemir V. Not all intelligence is artificial: data science, automation, and AI meet HI. Omi A J Integr Biol. 2019;23(2):67–9. https://doi.org/10.1089/omi.2019.0003.
111. De Choudhury M, Kiciman E. Integrating artificial and human intelligence in complex, sensitive problem domains: experiences from mental health. AI Mag. 2018;39(3):69–80 Retrieved from: http://kiciman.org/wp-content/uploads/2018/10/AIMag_IntegratingAIandHumanIntelligence_Fall2018.pdf.