Psychotherapy Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tpsr20

A scoping review of machine learning in psychotherapy research

Katie Aafjes-van Doorn , Céline Kamsteeg , Jordan Bate & Marc Aafjes

Psychotherapy Research, 2020
https://doi.org/10.1080/10503307.2020.1808729

EMPIRICAL PAPER

A scoping review of machine learning in psychotherapy research

KATIE AAFJES-VAN DOORN[1], CÉLINE KAMSTEEG[2], JORDAN BATE[1], & MARC AAFJES[2]

[1]*Ferkauf Graduate School of Psychology, Yeshiva University, Bronx, NY, USA &* [2]*Deliberate.ai, New York, NY, USA*

(Received 31 December 2019; revised 2 August 2020; accepted 3 August 2020)

Abstract

Machine learning (ML) offers robust statistical and probabilistic techniques that can help to make sense of large amounts of data. This scoping review paper aims to broadly explore the nature of research activity using ML in the context of psychological talk therapies, highlighting the scope of current methods and considerations for clinical practice and directions for future research. Using a systematic search methodology, fifty-one studies were identified. A narrative synthesis indicates two types of studies, those who developed and tested an ML model (k=44), and those who reported on the feasibility of a particular treatment tool that uses an ML algorithm (k=7). Most model development studies used supervised learning techniques to classify or predict labeled treatment process or outcome data, whereas others used unsupervised techniques to identify clusters in the unlabeled patient or treatment data. Overall, the current applications of ML in psychotherapy research demonstrated a range of possible benefits for indications of treatment process, adherence, therapist skills and treatment response prediction, as well as ways to accelerate research through automated behavioral or linguistic process coding. Given the novelty and potential of this research field, these proof-of-concept studies are encouraging, however, do not necessarily translate to improved clinical practice (yet).

Keywords: machine learning; psychotherapy; scoping review; big data; artificial intelligence

Clinical or methodological significance of this article: Machine learning (ML) offers robust statistical and probabilistic techniques that can help to make sense of big data. ML has recently gained popularity in fields such as psychiatric diagnoses and prognosis and pharmacological treatments. This review paper aims to broadly explore the nature of research activity using ML in the context of psychological talking therapies, highlighting the scope of current methods and considerations for clinical practice and directions for future research.

Machine learning (ML) lies at the core of artificial intelligence (AI) and data science, and at the intersection of computer science and statistics. Unlike most other computational strategies that involve a priori programing of fixed solutions (i.e. expert systems), ML provides computer systems the ability to automatically learn (i.e. self-learning algorithms) and improve from experience, in order to maximize accuracy (Jordan & Mitchell, 2015). In that sense, the goal of psychotherapists bears remarkable similarity to the goal of ML. Both psychotherapists and ML algorithms seek to accumulate knowledge from previous patients (datapoints) and translate it to a new patient's case, which may well be unique.

Machine learning involves the use of advanced statistical and probabilistic techniques to enable speedy and scalable analysis of complex or "noisy" data (e.g., non-linear, high-dimensional interactions). It thus offers new tools to tackle problems for which traditional statistical approaches are not well-suited (Bi et al., 2019). There is often no clear boundary between ML and statistical approaches (Bi et al., 2019). Indeed, whether a given methodology is considered ML or statistical often reflects its history as much as genuine differences, and many algorithms (e.g., least absolute shrinkage and selection operator, stepwise regression) may or may not be considered ML. Despite methodological

Correspondence concerning this article should be addressed to Katie Aafjes-van Doorn Ferkauf Graduate School of Psychology, Yeshiva University, Rousso Building, room 123, 1165 Morris Park Avenue, Bronx, NY 10461, USA. Email: katie.aafjes@yu.edu

similarities with traditional statistics, ML is philosophically and practically distinguishable in that, unlike traditional approaches that produce models that explain relationships between variables, ML emphasizes predictive accuracy (Bi et al., 2019). Although traditional statistical approaches are good for explaining, because they focus on goodness of fit based on a specific sample, these traditional models reduce generalizability that is needed for accurate predictions in other samples. Instead of testing for statistical significance, ML assesses performance of a model; that is, how accurate the model is in producing correct predictions or decisions when applied to a new dataset. Machine learning approaches are inherently suitable for use with "noisy", high dimensional (many variables) data (Barrett & Langdon, 2006), such as the large amount of verbal and nonverbal data in patient-therapist interactions. Machine learning enables patterns in data to be more readily and accurately identified, and more accurate predictions to be made from data sources (e.g. more accurate diagnosis and prognosis; Jordan & Mitchell, 2015), and has been found to be instrumental in identifying predictors and moderators where few consistent findings could be reached using traditional methods (e.g., Cohen & DeRubeis, 2018; Zilcha-Mano et al., 2018).

Predictive accuracy and validation are particularly important in determining whether and which ML algorithms are clinically useful, given the potential cost of an erroneous prediction (e.g., treatment failure, dropout and/or increase in symptoms).

There are two main types of approaches to the "learning" component of ML, each with a wide range of different algorithms: supervised learning and unsupervised learning (Graham et al., 2019). In supervised learning, a predictive model is developed based on both input and output data. The dependent variable values are known for each observation (i.e. specified outcome values) and are called "labeled data." (Bi et al., 2019).[1] Data with known labels are used to train a model that can predict the label for new unlabeled data (e.g., diagnosis of new patients in a clinic). Depending on the nature of the outcome variable, labels can be used for classification (prediction of categorical outcomes; e.g. diagnosis or not) or regression (prediction of continuous outcomes; e.g. along a spectrum of severity).

In contrast, in unsupervised learning, data is grouped and interpreted based on the input data only. The algorithm attempts to identify natural relationships within the data without reference to any outcome. Unsupervised learning thus uses unlabeled data to provide new insights, by representing data using less features (i.e. dimensionality reduction of continuous data) or clustering data in unspecified ways based on the data itself (Bi et al., 2019).

Besides the learning algorithm of a neural network that can be either supervised or unsupervised (depending on whether the desired output is already known), most other ML techniques cannot be used interchangeably. Commonly used supervised ML techniques include Support Vector Machines (SVM), K-Nearest Neighbors, Naïve Bayes, Regression techniques, Decision Trees, Random Forest, Hidden Markov Models, and Linear Discriminant Analysis. Commonly. Frequently used unsupervised ML techniques include K-Means Clustering, Hierarchical Clustering, and Principle Component Analysis (PCA). Another commonly used subtype of unsupervised learning involves deep learning algorithms that learn directly from raw data without human guidance, providing the benefit of discovering latent relationships in high dimensional data (LeCun et al., 2015).

Claims of the effectiveness of ML algorithms often detail the quality of the algorithm using several performance measurements. Although there is a lot of debate about which model evaluation metric is best or most appropriate (Handelman et al., 2019; Hernández-Orallo et al., 2012), the most commonly used evaluation metrics include, for example, accuracy, precision, F_1, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC[2]; see Graham et al., 2019).

In an attempt to improve the ability of models to generalize to new data that will be collected in the future, methods of internal cross-validation are commonly applied. Cross-validation is a technique to evaluate the performance of ML models in which researchers separate the available dataset into two parts; a training subset and a testing subset. Researchers then train a ML model on the training subset, and evaluate the resulting model on the complementary testing subset, using the evaluation metrics described above (Sammut & Webb, 2017; Tai et al., 2019). To reduce variability, many rounds of cross-validation are performed using many different partitions of the available dataset and then an average of the results is taken. These predictions may then be compared with the expected results, for example human codes assigned to the data or patient-reported outcome measures.

Different types of cross-validation can be applied (Sammut & Webb, 2017). For example, leave-*p*-out cross-validation (LpO CV) or leave-*one*-out cross-validation (LOOCV) are methods which learn and test on all possible ways to divide the original sample into a training and a validation set. In other cross-validation methods, such as *k*-fold cross-validation (10-fold is commonly used), and the holdout method, data points are randomly assigned to the training set and the test set. These types of cross-validation are referred to as internal validation, because the test set is still a subsample of the original dataset, rather than an entirely new or separate sample or population.

In addition to these steps of internal cross-validation, the additional step of out-of-sample external validation, is generally seen as the gold standard because it tests the performance of the newly developed predictive model on an independently collected dataset, rather than a subset or given percentage of the initial dataset used for training and internal cross-validation (Sammut & Webb, 2017). External validation thus reduces the risk that conclusions are being generalized from over-optimistic sample-specific ML predictions. The generated model is "overfitted" when it produces accurate results in the initial dataset, but is too specific to that dataset (i.e. has not ignored the "noise" in the dataset and has instead used parameters to account for that noise). As a result of being overfitted, the model is not accurate in making predictions in a new dataset. Even when the fitted model does not have an excessive number of parameters, it is to be expected that the fitted relationship will appear to perform less well on a new data set than on the training data set (a phenomenon sometimes known as "shrinkage"). Validating the model in an external sample is thus important to assess the extent of shrinkage and generalizability of the generated ML model and avoid accepting overfitted models.

Clinical applications of machine learning. Diagnosis and prognosis of medical diseases are perhaps the oldest clinical utilizations of ML techniques (Warner et al., 1961) and remain common applications in the epidemiologic literature (Bi et al., 2019). More recently, ML has also led to significant advances in the mental health field.

Psychiatry. Machine learning is known to help prediction research in psychiatry, by identifying robust, reproducible and generalizable predictors of treatment response in psychiatry (Gillan & Whelan, 2017). In their scoping review of ML applications in mental health, Shatte et al. (2019) highlighted a range of benefits across the areas of diagnosis, treatment and support, research, and clinical administration. Shatte et al. (2019) identified over 190 studies that applied ML in the detection and diagnosis of mental disorders, and over 60 studies to predict the progression of mental health problems over time, as well as explore computerized support for these mental health problems (Shatte et al., 2019). Studies, for example, used electronic health records, mood rating scales, brain imaging data, smart phone monitoring systems, and social media platforms to predict, classify, or subgroup mental health illnesses including depression, schizophrenia, and suicide ideation and attempts (also see Graham et al., 2019 for a selective review of 26 studies on AI in mental health). Similarly, Tai et al. (2019) concluded that ML can be used in unison with psychiatry by analyzing the multi-dimensional, multi-level disease models (i.e. dynamic interactions between molecular, cellular, and circuitry-based domains present in mental illness models). It is thought that ML will make it possible to help mental health practitioners re-define mental illnesses more objectively than currently done in the DSM-5 (Tai et al., 2019), and identify these illnesses at an earlier or prodromal stage when interventions may be more effective (Graham et al., 2019). Thus, in addition to disease-model refinement, ML may benefit psychiatry by characterizing those at risk, and personalizing and discovering pharmacological therapeutics (Tai et al., 2019).

Although reviews like those conducted by Shatte et al. (2019) and Tai et al. (2019) applied systematic literature search methods, they only searched for general keywords (e.g., "big data", "machine learning", "psychiatry" & "mental health"), and did not search for or report on psychotherapy. It is therefore not surprising that almost none of the identified studies in these reviews reported on ML in psychotherapy; only 8 out of the 300 studies identified by Shatte et al. (2019) and none of the 53 identified studies by Tai et al. (2019), examined psychotherapy. Arguably, most psychotherapists do not necessarily focus on diagnostic assessments or pharmacological treatment, and ML might be useful to them in different ways. Moreover, many of these ML applications in psychiatry are based on neuroimaging data, clinical notes, or electronic medical notes rather than the dyadic interactions between the clinician and patient that is prominent in psychotherapy.

Psychotherapy. Although several reviews have explored ML in mental health; no review has explored the breadth of ML applications in psychotherapy specifically. Some researchers have discussed the possible applications of ML in psychology (e.g., Yarkoni & Westfall, 2017) or clinical psychology more broadly (Dwyer et al., 2018). Based on their selective review of psychiatry research studies, Dwyer et al. (2018), for example, argue that the problems of translational clinical psychology and psychiatry that can be optimally addressed with ML fall into four main categories: diagnosis, prognosis, treatment prediction, and the detection and monitoring of potential biomarkers.

Within psychotherapy, technology enhanced human interaction, including ML, is likely to have a significant impact on (1) mechanism and process, (2) training and feedback, and (3) technology-mediated treatment modalities (Imel et al., 2017). In sum, the above-mentioned reviews on ML in psychiatry and mental health, and the examples reported in the psychology and psychotherapy literature highlight the potential relevance of ML in the field of psychotherapy and suggest that a comprehensive overview of ML applications in psychotherapy research is warranted.

Aims

Our overall aim was to examine the current state of affairs of ML applications in psychotherapy research, providing a snapshot of the applications relevant to the work of psychotherapists and psychotherapy researchers. As the field of ML is advancing exponentially, and the use of ML in psychotherapy is relatively new, we chose to focus specifically on exploring broadly the nature of research activity, as per Arksey and O'Malley's (2005) first goal of scoping reviews. It is hoped that this scoping review will: 1) inform clinicians of the methods and applications of ML in the context of psychotherapy; 2) clarify the strength and weaknesses of these methods and considerations within psychotherapy research; and 3) highlight clinical implications and identify potential opportunities for further research.

Methods

Inclusion and Exclusion Criteria

Studies were included when they reported on empirical investigations published in the English language. This excluded clinical and theoretical papers, study protocols and methodological papers. We were interested in all formats of psychotherapy, including online, face-to-face, individual, couple and group therapy, brief interventions as well as long-term treatment, and clinical-interviewing. Similarly, no limit was set on the nature of the mental health problem or patient diagnosis treated, treatment setting, or therapy modality. Studies on alternative forms of therapy, such as art therapy (e.g. Kim, 2008), pharmacological treatment (Koutsouleris et al., 2016) or electro-compulsive therapy (Redlich et al., 2016), or non-therapy interactions, such as peer-to-peer support (Jaroszewski et al., 2019) and primary care visits (e.g. Park et al., 2019) were excluded. We included studies that used data derived from psychotherapy treatments, regardless of the nature of the predictor variables used as input data (e.g., therapist notes/electronic health records/patient or therapist self-report measures/observer or supervisor ratings/ physiological or neuropsychological measurements). Studies that predicted symptoms or diagnoses, without using psychotherapy-related data, were excluded (for example studies on GP or emergency room data; Geraci et al., 2017). We also excluded empirical papers based on data in treatment manuals/websites rather than patient samples (e.g. Xing et al., 2017). Other studies were excluded because they examined the use of ML in clinical training of medical students, rather than psychotherapy students (Yang et al., 2019).

Studies were included when they reported on the development or use of a ML method, algorithm, or ML-based applications. Given that there is not always a clear difference between ML and statistical approaches (Bi et al., 2019) and that many algorithms (e.g., least absolute shrinkage and selection operator, stepwise regression) may or may not be considered ML, we based our criteria on the authors' descriptions of their analyses: if they deemed it ML, the paper was included.

In line with common practice in psychotherapy research, this scoping review focused on studies published in peer reviewed journals, to ensure replicability of the search procedures. We thus excluded book chapters, dissertations and other gray literature, including symposium and conference proceedings (e.g., Chaoua et al., 2018; Flemotomos et al., 2019), publications in the computer science journal INTERSPEECH that do not undergo peer review (e.g., Gibson et al., 2016; Singla et al., 2018; Tseng et al., 2017), and ArXiv (e.g., Ali et al., 2019; Crangle et al., 2019), a repository of electronic preprints that offers moderation of posting, but does not involve full peer review.

Search Strategy

The systematic literature search was conducted using widely available academic search engines across disciplines (PsycINFO & PubMed) and includes all IEEE Explore and ACM Digital Libraries. The last search was conducted October 31, 2019. Search terms included variations on the terms for: (a) psychotherapy (psychothera*, therap*, psychiatr*, counsel*, session*, interview*, "clinical assessment") and (b) ML ("machine learning", "artificial intelligence", "expert system", robotic*, "neural network*", "computer science", "computer vision", "natural language processing", "deep learning") (c) human (patient*, client*, "mental health").

The search was conducted on titles, keywords and abstracts with "AND" entered into the database search to link different categories (a, b and c) of search terms. Truncation symbols (*) were used to search for all possible forms of a search term. This means that 189 (7 × 9×3) separate searches were conducted for all the variations of the terms for psychotherapy (a), machine learning (b), and human (c). Please see Appendix A (see supplementary material) for the exact string of 189 search combinations. Forward reference searching, i.e., examining the references that cite these articles, as well as backward reference searching, i.e., reviewing the references that were cited in these articles, were applied to identify further studies that met the inclusion criteria.

Data Extraction and Analysis Plan

To conduct a systematic and accurate extraction of data, we developed a data extraction form, of which part is summarized in Table 1, and piloted it with all authors using five included studies. Building on study characteristics reported in other systematically conducted literature reviews (Graham et al., 2019; Shatte et al., 2019), the following data were extracted for each study: the first author, location, research question, study goal, sample and treatment characteristics, input data, type of ML algorithm and cross-validation, best performing results, conclusion and clinical implications. To capture the broad scope of studies, the data from the extraction form were synthesized using a narrative approach. A meta-analysis was deemed not appropriate given the humble aim of identifying research activity, and the wide range of study aims, ML techniques, and outcome measurements in the identified studies.

Results

Systematic Search Results

See Figure 1 for a schematic representation of the results at each stage of the systematic search procedure. The systematic search of 189 combinations of search terms identified a total of 5308 articles that included a search term from each category in their abstract or title. 906 articles were duplicates. Abstracts of 4402 articles were read by the second author to do an initial screening of eligibility for this scoping review. Of these 4402 articles, 4190 were excluded and 212 appeared to meet the inclusion criteria and were read in full by three of the authors. After reviewing the full text, 176 articles were excluded because they did not report on empirical findings, were not peer reviewed, did not involve psychotherapy or because they did not include ML approaches. A total of 36 relevant studies were identified. Forward and backward searches of references of these 36 studies were conducted and reviewed by the same authors and resulted in an additional 15 relevant studies that met the inclusion criteria. This resulted in a total sample of 51 studies that met inclusion criteria according to all four authors (see Appendix B (see supplementary material) for the full list of references of the reviewed studies). The selected 51 studies were reviewed in full to achieve consensus on the identified study characteristics.

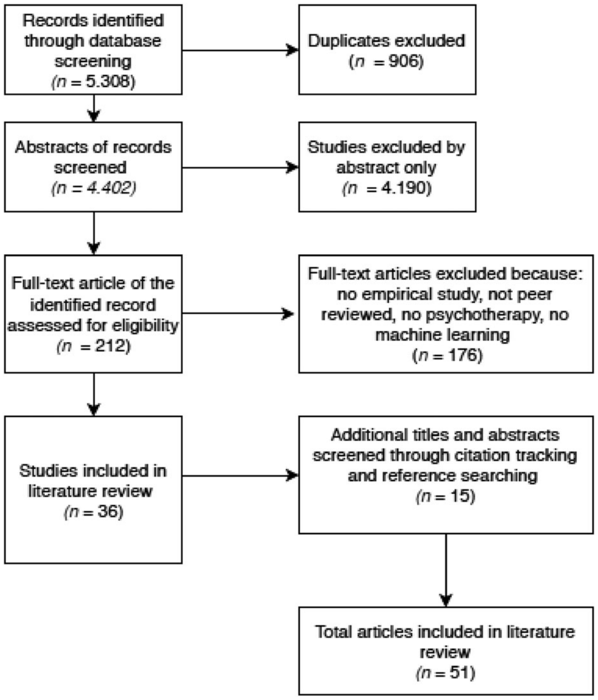
Records identified through database screening (n = 5.308)
Duplicates excluded (n = 906)
Abstracts of records screened (n = 4.402)
Studies excluded by abstract only (n = 4.190)
Full-text article of the identified record assessed for eligibility (n = 212)
Full-text articles excluded because: no empirical study, not peer reviewed, no psychotherapy, no machine learning (n = 176)
Studies included in literature review (n = 36)
Additional titles and abstracts screened through citation tracking and reference searching (n = 15)
Total articles included in literature review (n = 51)

Figure 1. Flow chart of systematic search procedures.

Study Characteristics

When we examined these 51 studies, we identified a subgroup of seven studies that reported on evaluations of treatment tools that included a ML element, rather than using any ML methodologies themselves. These implementation studies are indicated with an * in Appendix B (see supplementary material). More specifically, one study reported on the feasibility of a ML based training tool for therapists (ClientBot; Tanana et al., 2019), whereas Watts et al. (2014) evaluated the implementation of a ML tool to assess the use of evidence-based treatments from clinical notes. Krause et al. (2019) reported on using a fully automatized computer-based patient intervention for hazardous drinking

Table 1. Study Characteristics.

First Author	Year	Study Aim	Study Context	Sample size	Data Analyzed	Primary ML technique (S/U/C)	Cross-validation method (I/E)	Best algorithm performance on training sample and reported shrinkage
Althoff	2016	Predict outcome from text data	Non-profit organization delivering crisis intervention via text-message	408 counselors, 15,555 messages	During Session (continuous): Text messages between counselor and texter	Hidden Markov Model (HMM) (U), Logistic Regression Model (S) (C)	10-fold (I)	Accuracy: 68% AUC: 0.72
Atkins	2012	Replicate human ratings/codes/ judgments	Couple therapy corpus (UCLA/UW) from an RCT comparing two types of couple therapy	118 couples, 588 10-min interactions	Between Session - multiple points in treatment: Transcript of task/ exercise	Topic Modeling (U) & Sparse Logistic Regression Model (S) (C)	Leave-one-out (I)	Accuracy: 65% - 70%
Atkins	2014	Replicate human ratings/codes/ judgments	Addiction corpus based on 5 RCTs of MI in primary care settings and university students	148 sessions, 29,990 talk turns	During Session (continuous): Transcript of session	Labeled Topic Modeling (C)	10-fold (I)	AUC =.62 – 81 Average AUC=.72
Ball	2014	Predict outcome from pretreatment characteristics	Study of patients with generalized anxiety disorder or panic disorder receiving CBT	48 patients	Pre-Treatment: Neuroimaging (fMRI scans)	Random Forest Classification (S)	Hold-out (I)	Accuracy: 79% Sensitivity: 86% Specifi7city: 68%
Black	2013	Replicate human ratings/codes/ judgments	Couple therapy corpus (UCLA/UW) from an RCT comparing two types of couple therapy	117 couples, 569 10-min interactions	Between Session - multiple points in treatment: Acoustic speech and transcripts of task/ exercise	Support Vector Machine (SVM) (S) & Logistic Regression (S)	Leave-one-out (I)	Accuracy: 60.0-85.7% Average Accuracy = 70.15%
Bremer	2018	Predict outcome from pretreatment characteristics	E-Compared project, an RCT comparing blended internet and face-to-face treatment with TAU for depression	350 patients	Pre-Treatment: Patient self-report	Support Vector Machine (SVM) (S) & Regression Tree (S), LASSO (S), Ridge Regression (S)	Leave-one-out (I)	Regression Tree for outcomes, RMSE 0.0992 and Ridge Regression for costs, RSME 9187.78
Can	2016	Replicate human ratings/codes/ judgments & enhance behavioral coding	3 trials of motivational interviewing intervention studies	57 sessions	During Session (continuous): Session transcripts	Maximum Entropy Markov Modeling (S)	Leave-one-out (I)	F-score: 81.4 Sensitivity: 93% Specificity: 90% Precision: 73%
Connor	2007	Predict outcome from data	Comparison of 12-week CBT alone to CBT plus relapse prevention (acamprosate) for alcohol dependent patients	139 patients	Pre-Treatment: Patient self-report	Decision tree (S) & Bayesian Network (S)	Hold-out (I)	Accuracy: 77%

Ewbank	2019	Predict outcome from data	English IAPT program, patients receiving internet-enabled CBT for the treatment of a mental health disorder	14,899 patients, average 6 sessions per treatment	During Session (continuous): Session transcript Before Session: outcome measures Pre-Treatment and Post-Treatment change	bidirectional long short-term memory (BiLSTM) (S) / multivariable logistic regression modeling (S)	Leave-one-out (I)	NLP features model precision ranged from 52-100% Sensitivity ranged from 15-100% Specificity ranged from 79 −100%. Best Kappa of model-coder ranged from 0.24–1 for various features.
Foster	2019	Predict outcome from pretreatment characteristics	Treatment for Adolescents with Depression Study, an RCT comparing standard treatments for adolescents with major depressive disorder	439 adolescents	Pre-Treatment: Patient characteristics	Decision Tree (S) / Random forest blended with traditional parametric modeling (S)	NA	No specific evaluation of ML model
Gaut	2017	Replicate human ratings/codes/ judgments & enhance behavioral coding	Publicly-available psychotherapy corpus from Alexander Street press consisting of transcripts of patient-therapist conversations	1181 sessions, on average 250 talk-turns per session	During Session (continuous): Session transcripts	Labeled Latent Dirichlet Allocation (U)	10-fold (I)	AUC: 0.79
Gori	2010	Predict outcome from data	Patients of private-practice clinician over 12 months of treatment	150 patients	Pre-Treatment: Patient self-report	Artificial Neural Networks (ANN) (S)	Hold-out (I)	Accuracy: 90%
Hahn	2015	Predict outcome from neuroimaging	PANIC-NET, a German multicenter Mechanism of Action in CBT RCT study comparing CBT and behavior therapy for panic disorder/ agoraphobia	49 patients	Pre-Treatment: Neuroimagining (fMRI) during a differential fear-conditioning task	Gaussian Process Classifiers (GPC) (S)	Nested leave-one-out (I)	Accuracy: 82% Sensitivity: 92% Specificity: 72%
Hasan	2019	Replicate human ratings/codes/ judgments & enhance behavioral coding	Single MI sessions with African-American adolescents in treatment for weight-loss	37 sessions, 11,360 coded utterences	During Session (continuous): Transcript of session and audio recordings	Hidden Markov Model (HMM) (S)	NA	No specific evaluation of ML model. Study used ML methods to test hypotheses about what communication strategies lead to patient change. Previous studies reported on accuracy of models used, which was 75% for SVM and 87% for RNN.

(*Continued*)

Table 1. Continued.

First Author	Year	Study Aim	Study Context	Sample size	Data Analyzed	Primary ML technique (S/U/C)	Cross-validation method (I/E)	Best algorithm performance on training sample and reported shrinkage
Hoogendoorn	2017	Predict outcome from data	RCT on an internet-based guided self-help intervention for social anxiety disorder	69 patients	During Session (continuous): Transcript of session, multiple points during treatment	Decision Tree (S) / Logistic Regression (S) / Random Forest (S)	5-fold (I)	AUC:0.48-. 0.78 Precision: 77-88% Sensitivity: 35-73% F1: 49-89%
Idalski Carcone	2019	Replicate human ratings/codes/ judgments & enhance behavioral coding	Single MI sessions with African-American adolescents in treatment for weight-loss	37 sessions, 11,353 coded utterences	During Session (continuous): Transcript of session, each session	Support Vector Machine (SVM) (S) / Naive Bayes (S) / AdaBoost (S) / Random Forest (S) / DiscLDA (S) / Convolutional Neural Network (S)	10-fold (I) Out-of-sample test set: n=80 patient–provider interactions during routine HIV clinic visits (E)	F1-score: 68% Accuracy: 75.1% Out-of-sample test set: Accuracy: 72.0%. Kappa = .64
Imel	2015	Enhance linguistic coding	Publicly-available psychotherapy corpus from Alexander Street press and MI corpus from five randomized trials of MI for drug or alcohol problems	1,398 therapy sessions, 148 MI sessions, 1.2M individual words, 223K talk turns	During Session (continuous): Transcript of session	Unsupervised (U) and Supervised (S) Topic Modeling (C)	Hold-out (I)	Accuracy: 86.7%
Lenhard	2018	Predict outcome from patient characteristics	RCT of internet-delivered CBT for adolescents	61 patients	Post-treatment and 3-month Follow-Up: Patient-self report	Linear model with best subset predictor selection (S) / LASSO (S) / Random Forest (S) / Support Vector Machines (SVM) (S)	10-fold (I)	Accuracy: 83%
Lutz	2019	Predict outcome from data	Patients who began individual psychotherapy at a large university outpatient clinic, and either completed or dropped out	1234 patients, average 31 sessions per patient	Routine outcome measures: Patient self-report, therapist self-report, structured interviews Pre-treatment: Structured interviews	LASSO (S)	10-fold (I)	AUC: 0.67 Accuracy: 72.8% Sensitivity: 38.2% Specificity: 82.9% Kappa: 0.21

Mansson	2015	Predict outcome from neuroimaging	Study of internet-based CBT for social anxiety disorder	26 patients	Pre-Treatment: Neuroimaging (fMRI)	Support Vector Machine (SVM) (S)	Leave-one-out (I)	Accuracy: 91.7% Specificity: 100% Sensitivity: 83.3% AUC: 0.91
Mikus	2018	Predict outcome from data	E-COMPARED therapy project across five countries using the ICT4Depression/ MoodBuster platform	143 patients, 6428 data points (training), 6497 data points (testing)	Between Session: Daily EMA self-report	Recurrent Neural Networks (RNN), specifically Long-Short Term Memory (LSTM) Gated Recurrent Unit (GRU) (S)	Hold-out (50/50 splits; I)	Real mean standard error (RSME): 0.11
Modai	1996	Predict outcome from pre-treatment characteristics	Psychiatric inpatient unit	211 patients (training), 26 patients (testing)	Medical records/ patient information	Backpropagation Artificial Neural Network (ANN) (S)	Out-of-sample test set (E; n=26 patients)	Not reported
Nasir	2017	Predict treatment outcome from data	Couple therapy corpus (UCLA/UW) from an RCT comparing two types of couple therapy	134 couples, 3 sessions each, 458 10-minute interactions	Pre-Treatment, Post-Treatment. 2-year Follow-up: Video of problem solving interaction task outside of therapy session	Support Vector Machine (SVM) (S)	10-fold (I)	Accuracy: 79.6%
Nitti	2010	Enhance linguistic coding	Single case in private practice receiving supportive psychodynamic psychotherapy	1 patient, 43 sessions, 5,367 utterances	During Session (continuous): Session transcript	Competitive Neural Network (U)	NA	Not reported
Reggente	2018	Predict treatment outcome from neuroimaging data	Study of manualized ERP-based intensive CBT versus wait-list control	42 patients, 21 controls	Pre-Treatment and Post-Treatment: psychometric instruments & Neuroimaging (fMRI)	LASSO (S)	Random subsampling; Leave-10-out, k-fold (I)	Not reported
Rubel	2019	Predict treatment outcome from pretreatment characteristics	Psychotherapy patients in an outpatient clinic who received integrative CBT	115 therapists, 741 patients	Before Session: Patient self-report (outcome) After-Session: Patient self-report (alliance)	Boruta (based on Random Forest Classification) (S) / Nearest Neighbor (S)	Hold-out (I)	Correlations ranged from −.07 (300 NN=nearest neighbors) to .05 (30 NN); true error ranged 0.37 (1 NN) to 0.20 (100 NN and up). All above defined baseline of true error of 0.19.
Salomoni	2009	Predict treatment outcome from pretreatment characteristics	Patients referred to the Obsessive- Compulsive Spectrum Disorders Unit	130 patients	Psychologist-administered assessment	Artificial Neural Network (ANN) (S)	Random subsampling (I)	Accuracy: 93.3%

(*Continued*)

Table 1. Continued.

First Author	Year	Study Aim	Study Context	Sample size	Data Analyzed	Primary ML technique (S/U/C)	Cross-validation method (I/E)	Best algorithm performance on training sample and reported shrinkage
Schmitgen	2019	Predict treatment outcome from neuroimaging data	Previous studies of patients in two residential DBT programs	31 patients	Clinical characteristics, demographics, fMRI and sMRI	Random Forest (S)	10-fold (I)	Accuracy: 76.08% Sensitivity: 77% Specificity: 78%
Schultz	2018	Predict treatment outcome from neuroimaging data	Patients in treatment for major depressive disorder	21 patients, 20 controls	Clinical characteristics, demographics fMRI and sMRI	Support vector machines (SVM) (S)	Leave-one-out (I)	Accuracy: 88.9%
Seuchter	2004	Predict treatment outcome from text data	The A.N.I. study, a German multi-center study of the long-term treatment of schizophrenia	364 patients	Between Sessions: Session notes	Generalized Estimating Equations (GEE) (S) and Artificial Neural Networks (ANN) (S)	Random subsampling (I)	AUC: 0.66
Shiner	2013	Mimic human raters in classifying linguistic categories	6 VHA outpatient PTSD clinics	1924 patients, 84,561 clinical notes/ administrative data	Between Sessions: Administrative data and session notes	automated retrieval console (ARC), not further defined (S)	10-fold (I)	Sensitivity: 91% Precision: 100% F-measure: 95%
Sundermann	2017	Predict treatment outcome from neuroimaging	PANIC-NET, a German multicenter Mechanism of Action in CBT RCT study comparing CBT and behavior therapy for panic disorder/ agoraphobia	59 patients	Pre-Treatment and Post-Treatment: Neuroimaging (fMRI parametric maps)	Support Vector Machines (SVM (S))/ Multivariate pattern analysis (S)	Leave-one-out (I)	Accuracy: 66.70%
Symons	2019	Predict treatment outcome from data	Patients who sought CBT-based drug and alcohol treatment at a public hospital	830 patients	Pre-Treatment: Assessment data Each Session: breathalyzer and blood biomarkers for alcohol use	Fuzzy Unordered Rule Induction Algorithm (FURIA) (S), Bayesian network (S), various decision tree models (S)	10-fold (I) Out-of-sample test set (E; n=50)	Bayesian Network Accuracy: 62.79% Out-of-sample test set: FURIA: Accuracy of: 74% AUC: 0.49 Sensitivity: 31% Specificity: 89% Bayesian Network: Accuracy: 56% Sensitivity: 8% Specificity: 73%

Tanana	2016	Replicate human ratings/codes/ judgments	6 MI clinical trials	341 sessions, approx 1.7 million words, 175,000 utterances, 79,000 talk turns	During Session (continuous): Transcript of session	Discrete Sentence Feature (DSF) (S) and Recursive Neural Network (RNN) (S)	10-fold & hold-out (I)	Kappas for utterance codes: 0.00 - >0.50 Session level ICCs: .00-1.00
Tolmeijer	2018	Predict treatment outcome from neuroimaging data	Case-control cohort study of TAU or CBT for psychosis on top of usual care	38 patients	Pre-treatment: fMRI Pre-Treatment and Post-Treatment: self-report outcome measures	Multivariate pattern analysis (S) / multiple kernel learning (S)	Nested k-fold (I)	r=0.63, *p*=0.003
Tseng	2019	Replicate human ratings/codes/ judgments & enhance behavioral coding	Couple therapy corpus (UCLA/UW) from an RCT comparing two types of couple therapy	134 couples	During Session (continuous): Transcripts of interaction task	Nearest Neighbor / Neural Network (Long-Short Term Memory (LSTM)) (C)	Leave-one-out (I)	Accuracy: 78.77%
Tymofiyeva	2019	Predict outcome from pretreatment characteristics	Adolescent patients in CBT treatment from various psychiatric clinics	30 patients	Pre-Treatment: Neuroimaging (fMRI)	Decision Tree (J48 pruned tree classifier) (S)	10-fold (I)	Accuracy: 83%
Villmann	2008	Enhance coding of physiological data	Single-case study of manualized panic-focused psychodynamic psychotherapy	1 patient, 37 sessions	During session (continuous): Physiological data	Neural Network (growing self-organizing map (GSOM)) (U)	NA	Not reported
Wahle	2016	Predict outcome from data	Clinical pilot study of the smartphone app Mobile Sensing and Support (MOSS)	126 subjects	Between Session: Smartphone sensor data	Support Vector Machines (SVMs) (S) with a Radial Basis Function (RBF) kernel and Random Forest Classifier (RFC) (S)	Leave-one-out (I)	Accuracy: 61.5% Sensitivity: 62.3 Specificity: 60.8
Wallert	2018	Predict outcome from data	iCBT trial Uppsala University (Sweden) Psychosocial Care Programme (U-CARE) Heart study	90 patients	Pre-Treatment: self-report During Session; Adherence (completion of 2 homework assignments)	Random Forest Model (binary classifier) (S)	3×10–fold cross-validated recursive feature elimination (RFE) resampling (I)	Accuracy: 64%

(*Continued*)

Table 1. Continued.

First Author	Year	Study Aim	Study Context	Sample size	Data Analyzed	Primary ML technique (S/U/C)	Cross-validation method (I/E)	Best algorithm performance on training sample and reported shrinkage
Xiao	2016	Replicate human ratings/codes/ judgments & enhance behavioral coding	"TOPICS" corpus of 5 MI studies; & "General Psychotherapy" corpus of psychotherapy sessions in MI and other treatment types; "CTT" corpus of MI sessions from a therapist training study of Context Tailored Training	133 therapists, 826 sessions	During Session (continuous): Transcript of session and audio recording	Support Vector Machine (SVM) (S)	Leave-one-out (I)	Accuracy: 86%
Xiao	2015a	Replicate human ratings/codes/ judgments to automate detection of head movement	Couple therapy corpus (UCLA/UW) from an RCT comparing two types of couple therapy	574 10-min sessions	During Session (continuous): Video of therapy exercises (not sessions)	Gaussian Mixture Model based on Latent Dirichlet Allocation (LDA) model (U)	Leave-one-out (I)	Accuracy: 60-70%
Xiao	2015b	Replicate human ratings/codes/ judgments & enhance behavioral coding	"CTT" corpus of MI sessions from a therapist training study of Context Tailored Training; "General Psychotherapy" corpus of psychotherapy sessions in MI and other treatment types; selected data from "TOPICS" corpus of 5 MI studies	200 sessions	During Session (continuous): Transcript of sessions and audio recording	Support Vector Machine (SVM) (S)	Leave-one-out (I)	Accuracy: 85% Sensitivity: 96.7% Precision: 81.8% F-score: 88.6%
Zilcha-Mano	2019	Predict sudden gains from OQ data	RCT of five feedback conditions, conducted in an outpatient mental health clinic	28 therapists, 547 patients, 3174 sessions	Between Sessions: Self-reported outcome measurements & calculation of 'sudden gains' after each session	Classification & Regression Trees (S)	NA	Not reported - non-significant findings

Note. S = supervised, U = unsupervised, C = combination; I = internal, E = external; AUC = Area under the curve.

and depressiveness, and Burns et al. (2011) reported on Mobilize, a mobile phone sensing and ecological momentary assessment tool for depression. Similarly, another study reported using ML methods in a chatbot called "Wysa", that uses several evidence-based therapies (e.g. cognitive behavioral therapy, behavioral reinforcement, and mindfulness) to target symptoms of depression (Inkster et al., 2018). The other two studies reported on the therapists' experience of receiving automatized performance-based feedback, based on an ML algorithm (Hirsch et al., 2018; Imel et al., 2019). These seven studies imply that ML methods were used in the development of a particular tool, however, these studies were designed to evaluate a treatment/training more broadly, and did not report details on ML algorithms, training and testing procedures. In the subsequent narrative analysis, we will focus on the 44 studies that reported on ML model development (see Table 1 for a summary of the main study characteristics). The full data extraction form with all study characteristics is available from the authors upon request.

Publications. A narrative synthesis of the ML activity in the context of psychotherapy, indicated the emerging nature of this field, with most studies being published in recent years. Publication dates ranged from 1996 to 2019, however most articles are very recent. There is an 8-year gap between the first 1996 article (Modai et al., 1996) and the next study (Seuchter et al., 2004), and publications have accelerated recently with seven papers published in 2018 and 11 in 2019 (to date). The majority of studies were from the USA (*k* = 19) followed by Germany (*k* = 9) and the remainder (*k* = 16) split amongst several other countries in Europe, Australia and Asia.

Research questions and study aims. Most studies were initial proof of concepts to develop and test a ML algorithm for several purposes: a) predicting the response of a patient to a certain intervention/therapy (*k* = 27), either with regard to some outcome measure or completion/drop-out; b) automated behavioral coding (*k* = 12); and c) an analysis of the therapy process (*k* = 7).[3] Remaining studies focused on data outside psychotherapy sessions, such as identifying usage of certain types of therapy modality from clinical notes (Shiner et al., 2013).

Study design and context. Of the 44 studies, 32 were randomized controlled studies, and 11 took place in a naturalistic setting, and one study used training recordings. The ML model was applied to cross-sectional data in 14 studies where the aim was to replicate human codes assigned to data or enhance physiological or linguistic codings, and 30 studies used ML to predict outcomes in longitudinal data.

Sample and treatment characteristics. Sample sizes ranged from single case studies (e.g. Villmann et al., 2008) to larger scale including 14,899 patients, with 90,000 transcripts (approx. 200M words; Ewbank et al., 2019). It is important to highlight that most study samples were relatively modest, with just 14 studies having a sample of 200 or more patients. The studied samples mostly consisted of individual adult treatment sessions but five studies used data from couple therapy, and four studies included a pediatric or adolescent population.

A variety of treatment modalities were covered and many studies included multiple modalities within the same analyzed sample. The largest included category was CBT (k = 12) followed by Motivational Interviewing (*k* = 8), integrative (*k* = 6), internet-based CBT (*k* = 5) and psychodynamic (*k* = 2) modalities. Other modalities included Cognitive Processing Therapy, Exposure and Dialectical Behavior Therapy. In many studies, treatment approach was undefined (*k* = 16) and six studies included (partial) medicated populations.

Many studies covered samples including patients with different diagnoses and a wide range of diagnoses was covered. The largest categories were depression (*k* = 10), panic & anxiety (*k* = 9), addiction (*k* = 9), OCD (k = 3), Schizophrenia (*k* = 3), Borderline (*k* = 2), PTSD (*k* = 1) and Bipolar disorder (*k* = 1). A large number of studies also included generalized or unspecified problems (*k* = 16).

Type of data. Most studies use transcripts of sessions (*k* = 15) with/or without other textual information from medical records (*k* = 2), demographics (*k* = 2) or session notes (*k* = 2). In total 16 studies applied text or Natural Language Processing (NLP) analyses.[4] Natural Language Processing is a specific type of AI that may be used to transform raw texts (i.e. clinical notes, session transcripts) into more useful labeled data (Pace et al., 2016; Weusthoff et al., 2018), and might be used prior to performing ML algorithms.

Outcome data in many cases consisted of questionnaires completed by clinicians (*k* = 7) or through patient self-report measures (*k* = 12). There was also a distinct set of studies published in medical journals that combine a variety of metrics with neuroimaging data (*k* = 7), primarily fMRI (all of these studies focus on psychotherapy treatment outcome prediction).

Fourteen studies utilized session transcripts. But of note is the small number of studies analyzing other observable metrics during or in between therapy sessions such as audio acoustics ($k = 4$), video (Nasir et al., 2017), biometrics (Villmann et al., 2008) or ambient smartphone data (Wahle et al., 2016). It seems this could be a fruitful avenue for further research as well as shown by the prediction value of acoustics analysis in couples therapy (Nasir et al., 2017) and opportunities identified in Shatte et al. (2019) or by multimodal studies (e.g. Dibeklioğlu et al., 2017).

Type of ML method and validation. Most studies reported on supervised ML applications ($k = 35$). Four studies used unsupervised learning, and five studies reported using both supervised and unsupervised ML approaches. Overall, studies applied a very broad range of ML techniques; artificial neural networks ($k = 11$) and support vector machines ($k = 11$) were most frequently used.

The most commonly used types of cross-validation were leave-one-out ($k = 14$), 10-fold ($k = 13$), and hold-out ($k = 7$). Six studies used other methods and four did not report cross-validation. While many studies described the technical approaches to their cross-validation, limited rationale was provided for their choice of approach. For example, 10-fold was used seemingly as a standard of reporting, much like an arbitrary *p*-value.

Reported best performing algorithm. There was considerable heterogeneity in the nature of how the results were reported across studies; some studies did not report any algorithm performances (k=7), whereas others provided multiple metrics for each ML model that was tested. Also, 14 studies only reported an accuracy measure and did not provide a confusion matrix or another way to understand attributes, such as false positives or negatives. The most commonly reported evaluation metrics were accuracy ($k = 25$), sensitivity ($k = 12$), specificity ($k = 9$), AUC ($k = 8$), precision ($k = 5$), F1 score ($k = 5$), and/or kappa ($k = 4$). The accuracy rates of the best performing algorithms ranged from 60-99.2%, Specificity ranged from 68% to 100%. Sensitivity (i.e. recall) ranged from 8% to 100%. Precision ranged from 52% to 100%. The AUC ranged from .49 to .91. F_1 was only reported in 5 studies and ranged from 68% to 95%.

Of the 12 studies that examined whether ML models could be used to predict behavioral or observational codes (i.e. ratings/labels) assigned by human raters or experts, seven reported the accuracy of the best algorithmic performance, ranging from 60% (Xiao, Imel, et al., 2015) to 86% (Xiao et al., 2016). One study reported sensitivity (62%; Atkins et al., 2014), two studies reported specificity, ranging from 72% (Atkins et al., 2014) to 90% (Can et al., 2016). Two studies reported precision, which ranged between 73% (Can et al., 2016) and 81.8% (Xiao, Georgiou, et al., 2015). Three studies reported F_1-scores, ranging from 68% (Idalski Carcone et al., 2019) to 88.6% (Xiao, Georgiou, et al., 2015). Two studies reported AUC ranging from .716 (Atkins et al., 2012) to .789 (Gaut et al., 2017). Two of the twelve studies reported reliability of the algorithms with human coders in terms of Cohen’s kappa (Idalski Carcone et al., 2019; Tanana et al., 2016).

Of the five studies that examined ML models to identify characteristics of sessions (transcripts or texts) that predict outcomes (end-of-treatment or within session), two studies reported accuracy, ranging between 68% (Altoff et al., 2016) and 79.6% (Nasir et al., 2017), two studies reported precision, ranging from 52% to 100% (specific features reported by Ewbank et al., 2019). Two studies reported an AUC, ranging between .716 (Althoff et al., 2016) and .780 (Hoogendoorn et al., 2017). Only Ewbank and colleagues (2019) reported on sensitivity, ranging from 15-100%, and specificity ranging from 79. −100%.

Thirteen studies used ML models to predict treatment outcome based on pre-treatment or questionnaire/intake data. Six studies reported accuracy, ranging from 62.79% (Symons et al., 2019) to 99.2% in the training set reported by Gori et al. (2010). Four of these studies did not report best algorithm evaluation metrics or reported true error (.20; Rubel et al., 2019) or AUC (.66; Seuchter et al., 2004). Lutz and colleagues (2019) were the most comprehensive and reported accuracy (72.8%), sensitivity (38.2%), specificity (82.9%).

Of the nine studies that predicted treatment outcome based on neuro-imaging data, the reported accuracy ranged from 63% (Tolmeijer et al., 2018) to 91.7% (Månsson et al., 2015). Sensitivity ranged from 77% (Schmitgen et al., 2019) to 92% (Hahn et al., 2015). Specificity ranged from 68% (Ball et al., 2014) to 100% (Månsson et al., 2015). Only Månsson et al., (2015) reported the AUC of the best performing algorithm (.91). Reggente et al. (2018) did not report performance metrics.

Of the four studies that demonstrated the use of ML analytics for linguistic coding (Imel et al., 2015; Nitti et al., 2010) or coding of physiological data (Villmann et al., 2008), Imel and colleagues reported accuracy (86.7%) and Shiner et al. (2013) reported sensitivity (91%), precision (100%) and F_1

of 95%. The two studies that used ML models to predict treatment outcome based on ecological momentary assessment during treatment, reported an accuracy of 60% (Wahle et al., 2016) and a real mean standard error of .11 (Mikus et al., 2018).

Generalizability. Three studies tested predictions out-of-sample using external validation to determine the generalizability of the ML models. Idalski Carcone et al. (2019) utilized a new test set of 80 participant-provider interactions to validate the SVM model that performed best in the cross-validation study, and reported accuracy of 72.0%, with reliability comparable to human coders (kappa = . 639), and a shrinkage of accuracy from original to out-of-sample of 2.9%. Symons et al. (2019) used an external test set of 50 patients and reported accuracy of the 10 best performing models from the cross-validation. Although the Bayesian Network model had performed best on the cross-validation (with a reported accuracy of 62.79%), when applied to the external test set the accuracy of this model reduced to 56%. The best performing model on the external test set (the FURIA model), had not been the best performing model on the cross-validation, and had a reported accuracy of 74% in the external validation sample. Modai and colleagues (1996) used an external validation test set of 26 patients to examine the accuracy of the Adaptive Resonance Theory Neural Network (ART) algorithm to make treatment decisions. They did not report on the accuracy or other metrics of this ML model specifically, but compared the efficacy of treatments suggested by the clinician to the efficacy of treatments suggested by the ART algorithm and concluded that both methods were similarly effective.

Software and open source libraries. Many studies did not mention which statistical tools were used for analysis ($k = 16$), but for those that did, most used a variety of software packages in R and Matlab (but also SPSS, Stata, Statistica, SAS, and Weka). The studies reported on using standard libraries available for data preparation (e.g. missing variables), a variety of typical ML models and NLP analyses (such as topic modeling) included in their standard packages like R or Matlab. Some specialized tools were used for study-specific analysis and all of those were available as libraries for R or as standalone packages (mostly written in Python). No specific software seems to have been developed for the reviewed studies. Only one study (Lutz et al., 2019) has shared its resulting ML model (and training data) publicly. A few studies used open-source treatment data that was widely available (e.g., Gaut et al., 2017; Imel et al, 2015).

Study conclusions. Most of the 44 reviewed studies concluded that ML models were effective in predicting the target, whether it was human codes used to label data or treatment outcomes, and implied that the ML approach was more beneficial than previously applied traditional statistical approaches. However, as described above, the level of accuracy, sensitivity, or specificity that is considered to be acceptable varies depending on the aims of the study and the dataset. None of the studies explicitly compared the ML performance with that of more traditional statistical analyses. In the majority of published studies, the ML approach aided the researchers in answering their research question. Zilcha-Mano et al. (2019) provided a notable exception in that even when using the ML method, they were unable to identify patients most likely to show sudden gains in treatment.

Discussion

Summary

Building on recent literature on the potential of ML approaches in psychiatry, mental health, and psychology in general, this review aimed to systematically scope the empirical literature on ML applications in psychotherapy research. More specifically, we aimed to 1) inform clinicians in the methods and applications of ML in the context of psychotherapy; 2) clarify the strength and weaknesses of these methods and considerations within psychotherapy research; and 3) highlight clinical implications and identify potential opportunities for further research. Fifty-one studies were identified in a systematic search across peer-reviewed journals in mental health and computer science.

Ml Methods and Considerations in Psychotherapy

Sample sizes. It is generally accepted that ML algorithms require larger sample sizes than traditional statistical methods (Schwartz et al., 2020), but exactly how large remains unclear. Most ML studies in the field of psychotherapy so far appear to have ignored issues of proper sample size calculation and replicability/generalizability, implying that high-tech ML algorithms are immune to issues of low sample power. Some of the reviewed studies applied ML to relatively small numbers of participants and justified the use of ML based on the large variety of variables/datapoints for each participant. The concern in such situations is that models will be overfitted, that is, specific to that dataset, and will not be accurate in making predictions in a new dataset. Most of the reviewed studies appear to have glossed over questions such as; How large should the training sample be? How large should the validation sample be? Do we need separate calculations for each set, or to decide how to apportion a % of cases to a training set and a % of cases to a test set? This lack of explicit mention of sample size calculations might not be surprising given the lack of clarity on this in the ML field more broadly.

Although most agree that the ideal sample size needed for ML depends on the quality of data and the complexity of the model with regards to sample size (volume), multitude of data sources (variety), and how quickly it's accumulating and changing (velocity), the required minimum sample size in ML is a fertile ground of methodological discussion (see Balki et al., 2019). The general rule of thumb is that the amount of training data needed for a well performing model is 10x the number of parameters in the model (Caballero et al., 2006). Although no minimum sample size calculations were provided in the reviewed studies, they indeed appeared to report on sample sizes that were at least ten times the number of modeled variables. Notably, this ten-cases-per-predictor rule is not universally accepted, and is contested by many on statistical grounds. For instance, the sample size required to predict an event with a low base rate (e.g., suicide) will be considerably larger than that required to predict a more common event (e.g., reliable improvement of symptoms post-treatment). The 10-cases rule ignores base rate and effect size considerations, which are basic parameters in conventional sample size calculation strategies for classification problems (Hsieh, 1989). For supervised ML models, Sahiner et al. (2008) discuss sample size considerations in relation to resampling and cross-validation techniques. More recently, Riley and colleagues developed sample size calculation guidelines for multivariable prediction models (Riley et al., 2019a, Part I and Part II) that consider expected effect sizes, the distribution of predictors, the ratio of predictors-to-cases, and the expected prediction shrinkage factor for cross-validation designs, and the number of events in each category of categorical predictors (Riley et al., 2019b).

For unsupervised ML models (e.g., detecting the correct number of latent classes in latent profile analysis), adequate power is reliant on a very large degree of distinction between predictors, whereas sample size and the number of predictors appear to matter only when the degree of separation is lower (e.g., <0.80 instead of >1.0) (Tein et al., 2013). Similarly, in classification models, the relative size of training samples should be weighted appropriately to reflect the "complexity" of each class in order to avoid classification bias. This means that broadly defined classes with a high intra-class variability, should be trained on larger samples than more narrowly defined classes (Blamire, 1996).

Model performance. Overall the reported best performances of the ML models appeared promising. With future larger datasets, these accuracies would be expected to be higher, given that many ML algorithms increase in performance with larger datasets (Raudys & Jain, 1991). The authors of most reviewed studies implied that their ML models helped them to answer their respective research questions (except for the study conducted by Zilcha-Mano et al., 2019), but they did not directly compare the performance of ML methods with traditional statistics. The issue of which is better is widely debated in the literature outside of psychotherapy. Machine learning methods may be advantageous when working with large datasets with a greater number of predictors, particularly if those predictors and covariates are not normally distributed (Hsieh, 1989) and when there is a high signal-to-noise ratio (Christodoulou et al., 2019). However, a systematic review of 71 studies concluded that clinical prediction models trained using ML analyses did not significantly outperform those trained using logistic regression (Christodoulou et al., 2019). Whether this conclusion applies to psychotherapy data and classification problems is unknown, simply because current studies failed to compare ML algorithms versus parsimonious statistical models such as logistic or linear regressions using conventional backward elimination and no resampling techniques. In future studies ML models should be benchmarked against their simpler statistical counterparts, so that our field can advance in its appraisal of the relative costs and benefits of ML techniques.

Others argue that, rather than pitting methods against each other, both ML and traditional statistics approaches should be applied, as these might result in the identification of different variables (Cohen et al., 2020; Schwartz et al., 2020). Indeed, several of the reviewed studies reported on both traditional statistical analyses and ML models (e.g. Atkins et al., 2012; Seuchter et al., 2004) and explicitly stated that the numeral results were not directly comparable. Models that combine traditional statistics and ML algorithms appear promising in training samples, but have not yet been replicated in hold-out samples (Schwartz et al., 2020), and appear to have large shrinkage in external validation samples (Delgadillo & Gonzalez Salas Duhne, 2020).

Some reviewed studies reported varying degrees of accuracy and were not always explicitly clear regarding the meaning of resulting performance metrics. Although this might raise concerns over systematic overestimation and methodological inconsistencies, it is a commonly held misperception that a model needs to be highly accurate to be of clinical use. The evaluation metric and what is considered an acceptable level of accuracy depends on what the algorithm will be used for. For example, in a large data set that is also one in which errors would not be particularly costly, then a lower level of accuracy might be considered acceptable and beneficial because it would increase efficiency and thereby the sample size. Such an example in psychotherapy research would be the use of ML algorithms in process coding. These algorithms may make errors in coding but could vastly increase the amount of data that could be coded, thus outweighing the impact of errors made in coding. Similar to the application in psychotherapy research, a ML model with a relatively low accuracy might still be beneficial to clinical practice. Performance accuracy, for example, would be more clinically informative in comparison to clinical diagnostic accuracy (as opposed to simply relating these values to chance; Graham et al., 2019). If a model can outperform current clinical practice, then patient utility can be maximized at scale. Whilst far from perfect, an increase in prognostic certainty (e.g. from 50 to 65%) might be clinically meaningful (Cearns et al., 2019). Therefore, the absolute accuracy of a classifier should not serve as an indicator of clinical utility, but the relative increase in prognostication compared to current practice. However, in a different clinical context where treatment decisions are being made based on ML, errors may be quite problematic, and thus require a high level of accuracy.

Generalizability. Two things are important for researchers to understand about validation. First, both internal and external cross-validation approaches give expected out-of-data-set performance given the algorithm used, not an assessment of the particular fitted model. Secondly, the model performance depends on the representativeness of the sample used for training compared to the overall population. This means that, if the training and test set are highly correlated, cross-validation performance will be deceptively high, reflecting a bias of the composition of the training data. Few studies in our scoping review tested the algorithm on an independent sample (an issue also highlighted in the systematic review by Christodoulou et al., 2019). It is important, therefore, to test out-of-sample predictions, not just on a subsample of the original sample (internal cross-validation), but also on a different new sample (external validation) to confirm their clinical utility, particularly because methods for resampling have been found to produce different results (e.g., Sahiner et al., 2008).

Although these proof-of-concept studies, are encouraging, simply predicting certain symptoms or behaviors, does not (yet) necessarily translate to clinical practice. These reviewed studies have limitations pertaining to clinical validation and readiness for implementation in psychotherapy. As is true for any ML application, the size and quality of the data limit model performance (Graham et al., 2019). Moreover, most studies only tested the ML models within the same sample (training and test set) rather than with external validation, which limits the generalizability of the results. The predictive ability of studies is restricted to the features (e.g., clinical notes, session transcripts, biomarkers, human codes) used as input for the ML models, and the clinical efficacy of the features used to derive these models must be considered. It remains possible that the outputs of these algorithms are only valid under certain situations or for a certain group of people. Thus, in order to provide the most clinically relevant information, advanced ML algorithms should be trained based on a wide variety of naturalistic datasets.

Assessing big data confidentiality. The use of big data (defined as extremely large data sets of multiple rapidly changing variables that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions; Big Data, 2020) also requires a new approach to research practices. Although most will appreciate the importance of storing and sharing treatment data in a confidential way, few might be aware of the specific challenges inherent to big data. Protecting identities is not merely a matter of deleting or masking a person's name or other specific identifier. With the large sample sizes and range of data types and sources, identities can also be reconstructed by combining pieces of information, each of which would not be enough to identify a person but, combined, would allow individuals to be identified (Berman, 2013). Given the movement of archiving and distributing data more openly (e.g., the Open Science Framework: https://osf.io/), which frequently occurs with big data in computer science, additional care must be taken to ensure that individuals cannot be identified in unexpected ways (Chen & Wojcik, 2016). Currently, there appears to be a lack of guidance on development of ML applications, their clinical integration and training of psychotherapists, as well as a "gap" in ethical and regulatory frameworks (Fiske et al., 2019). Institutional Review Boards' may also have limited knowledge of emerging ML methods and applications, which makes risk assessment inconsistent.

Improving machine learning expertise. For those who are unfamiliar with the field of ML, the emerging research can be daunting, with a wide variation in the terms used and the metrics presented. Similar to the field of psychotherapy, where researchers and clinicians debate the relevance of significant, clinical and reliable change, there is a great deal of debate about which ML model evaluation metric is best (Handelman et al., 2019; Hernández-Orallo et al., 2012). Making sense of reported ML evaluation metrics is made even harder by the fact that different performance parameters often provide conflicting results and that the optimal ML algorithm also depends significantly on the composition of the dataset (Rácz et al., 2019).[5]

This means that sometimes it may be desirable to select a model with a lower classification accuracy and instead report precision, recall or F_1 because it has a greater predictive power on the problem. In general, one can try to train a model that achieves the best possible values for all metrics. However, in practice, a pragmatic application-oriented approach is more relevant (e.g. in predicting suicidal ideation one might want to have a model that minimizes false negatives over minimizing false positives). After all, the metric values only compare the models to each other, whereas how good a model really is can only be shown in practice.

A specific point to note is that terms such as dataset, sample and training subsample are often used interchangeably in way that can create confusion to the uniformed reader, or to those missing the broader study context. Part of this might stem from the merging of the fields of computer science and behavioral science. In behavioral science "sample" would be used to refer to a collection of data from a larger population, such as patients in psychotherapy. A computer scientist might simply refer to this sample as the dataset from which to draw training and test samples (which would more accurately be subsamples). This particularly can create confusion when reporting external validation results where the "training sample" might refer to the full dataset (sample) rather than a sub-sample typically used for internal cross-validation.

A general lack of working knowledge on ML algorithms, despite their substantial methodological overlap with statistical methods (Beam & Kohane, 2018) reduces the practical uptake of these techniques and increases the risk of misinterpreting data and misusing these methods; for example, to make overly optimistic claims about findings where out-of-sample predictions were not reported. The current lack of procedural evaluation guidelines leaves many non-expert clinicians and researchers in the field with no means to systematically evaluate the claims, maturity, and clinical readiness of a ML study (see also Cearns et al., 2019 on ML in psychiatry).

ML techniques may also be off-putting to those interested in theory, because ML models are thought to be black boxes where interpreting how a model works, or especially why a subject is classified, is difficult (Hart & Wyatt, 1990). Although some techniques can be applied to explain the prediction of classifiers (Ribeiro et al., 2016), interpreting multiple latent variables (e.g., in deep learning) is complicated, and more work is required to bridge the gap between ML in psychotherapy research and clinical care. Thus, it will be important for psychotherapy researchers to become better-versed in the ML methods and how to interpret this research literature. Based on the reviewed studies, it might be particularly useful to become familiar with software programs such as R and Python. There is a plethora of open-source statistical tools readily available which would make this expanding field easily accessible to many psychotherapy researchers. Accessible ML education and tool development is required to facilitate understanding and usage in the wider clinical research community. Besides formal education on ML in psychology graduate programs, it might also be helpful for psychotherapy researchers to attend (online and freely available) courses on ML.[6] When conducted with care for ethical considerations, ML research can become an essential complement to traditional psychotherapy research.

Clinical Implications & Future Research

The range of applications of ML models appears to address several important directions in the field. First, ML may help to bridge the gap between science and practice. It is important to highlight that none of the identified ML applications were developed to replace the therapist, but instead were designed to advance the therapists' skills and treatment outcome. In this way, ML might become part of evidence-based practice, as another source of valuable information, in addition to clinical intuition, patient's preferences and existing research evidence. Greater collaboration between researchers and clinicians (e.g., for the provision of training data sets, and for the feedback on the clinically usefulness of ML algorithms) will be needed to continue to advance the applications of ML in psychotherapy. ML methods provide an opportunity for multimodal analyses of patient and therapist moment-by-moment changes in word use, speech, body movements, and physiological states, that are not (yet) usually considered in clinical decision making.

ML methods are also aligned with the increasing interest in tailoring interventions to individual patients, pinpointing which treatments provide the optimal benefit for which patients at the right time (e.g., Persson, 2019). For example, ML approaches could help identify personalized process-outcome associations in psychotherapy (Rubel et al., 2019). There is a rapidly growing literature in this area, specifically, with recent demonstrations of ML techniques used to develop treatment selection models (CBT versus psychodynamic psychotherapy or person-centered counseling for depression; Cohen et al., 2018; Delgadillo & Gonzalez Salas Duhne, 2020; Schwartz et al., 2020). Analyzing "big data" on clinical outcomes across large combinations of different treatment approaches, crossed with the multitude of genetic, biomedical, behavioral, environmental, and demographic patient characteristics could help predict different responses to different treatments. Even if no moment to moment session recordings are available, existing baseline characteristics from electronic health record databases can provide valuable, real-world, practice-based evidence to support better models of predicting which patients might do well in treatment (e.g., Zilcha-Mano et al., 2019) or which treatments may be most effective for a particular patient (Lutz et al., 2019). In this way, ML offers a solution to analyzing idiographic research questions in big-data (Silberschatz, 2017).

Moreover, ML approaches might be very suitable for transtheoretical psychotherapy research. This might be especially relevant for process research. Once the infrastructure to automatically collect patient data and recording of psychotherapy sessions becomes more common, researchers may be able to more efficiently code theoretically identified process variables in large datasets. As Imel and colleagues point out, ML approaches may ultimately be used to predict response to therapy using statistical models that only rely on objective measures of the session process and that are not limited by our current theoretical understanding of the mechanisms of change (Imel et al., 2017). Multi-site collaborations that generate large practice-based datasets may allow for representative samples as well as external cross-validation of ML models.

Furthermore, the transparency of published algorithms, shared codes, and open source data (e.g., the couples therapy corpus and general psychotherapy corpus in the reviewed studies), which is common in the field of computer science, may help address the current replication crisis (Tajika et al., 2015), and the relative lack of funding in psychotherapy research. Given that, for example, National Institute of Mental Health (NIMH), IOM (Institute of Medicine), and the Affordable Care Act (see also National Institute of Health (NIH), Precision Medicine Initiative Working Group; Intille, 2016), explicitly support the assessment of treatment quality on a large scale, ML based psychotherapy research projects are likely to attract external funding from grants and agencies.

Limitations of This Review

First, with this review we aimed to provide a snapshot of the breath of research activity in an accessible and summarized format, while using a systematic search method. This review, could be seen as a scoping review, in that it provides a rigorous and transparent method for mapping complex areas of research, in terms of the volume, nature and study characteristics (Arksey & O'Malley, 2005). In line with the aims of a scoping review, we did not identify specific study designs in advance, and did not address specific research questions nor assessed the quality of included studies (Arksey & O'Malley, 2005). We acknowledge that this is one review method amongst many that might be beneficial to the field, and hope that psychotherapists, psychotherapy researchers, patients and policy makers will be well placed to build upon our reported findings.

Second, from writing this review, we have learned that identifying ML methods in published studies is not straightforward, especially given the wide variety of possible ML techniques and applications, and related technical jargon, and the lack of a clear boundary between ML and statistical approaches (Bi et al., 2019). It is possible that studies that used ML but did not explicitly state this by using any of the included search terms, were not identified in this scoping review. Similarly, the large number of studies identified through backward and forward reference searches, suggests that studies might not always use common ML terms in their title, abstract or keywords, even if they use ML techniques to analyze their data. Ironically, the systematic literature search process (which is currently based on automatic term recognition) could be advanced by using ML algorithms to help identify relevant studies, that might not have used specific pre-determined search terms but nevertheless meet the inclusion criteria. ML, for example, might in future offer the possibility of finding useful associations among disparate facts, leading to the discovery of new or unsuspected knowledge (Thomas et al., 2011).

Third, during the systematic search procedures, we identified several relevant empirical papers that reported on ML applications to psychotherapy, but that were not published in peer reviewed journals. For example, a conference proceeding by (Tanana et al., 2015) reported on sentiment analyses in 97,497 utterances from psychotherapy transcripts labeled by humans. Although this same research group published in peer-reviewed papers since, this exact study was not (yet) identified in the peer-reviewed literature. Another example of an identified relevant study not included in the current scoping review is the dissertation by Hasan (2019) on ML methods for the analysis of clinical conversation (automatic annotation, sequential analysis, & segmentation).

Moreover, restrictions in the search methodology may have resulted in relevant articles being missed. The field of ML is rapidly developing and can make any review seem obsolete within months (Bi et al., 2019), especially when adhering to the publication standard common in psychotherapy research, where the peer-review process of relevant studies can take months if not years. This means that, it is possible that new ML applications in psychotherapy research are available by the time this scoping review is available for publication, and that additional applications might have been reported at conferences or non-peer reviewed outlets but could not (yet) be identified in a search of the peer-reviewed literature. Although in the medical field, the peer-review process is held in high regard, and is deemed crucial in controlling the quality of statistics in publications (Bacchetti, 2002), in computer science the quick turn-around of reporting of scientific findings is often preferred, and write-up of for conferences proceedings and self-archiving in the arXiv repository is very common (personal communication with Professor David G. Lowe, November, 2019). It is thus likely that the inclusion of studies published in non-peer-reviewed outlets, such as ArXiv, IEEE conference proceedings or INTERSPEECH would have indicated an even broader scope of possible future opportunities.

Furthermore, this scoping review focused on the application of ML to psychotherapy specifically. However, as indicated in the introduction, ML might be usefully applied in broader aspects of the work psychotherapists do, such as diagnosis, assessment, and policy and training (see Shatte et al., 2019). A second major barrier in the translation of ML methods is that the techniques are difficult to understand and implement for many clinicians and clinical researchers. This review has attempted to provide an understanding of critical techniques, and there are some excellent textbooks that can assist researchers with statistical knowledge (e.g., James et al., 2013).

Conclusions

Most identified studies in this scoping review should be considered early proof-of-concept works demonstrating the potential of using ML algorithms to address psychotherapy questions, and which types of algorithms yield the best performance. Caution is necessary in order to avoid over-interpreting preliminary results. The use of ML in and of itself does not necessarily increase the chance of treatment success or improve clinical decision making, but by gathering diagnostic information, clarifying treatment processes, and reviewing therapist behavior, ML is poised to impact traditional approaches to delivering psychotherapy (Miner et al., 2019). Compared to traditional statistical methods, ML brings new possibilities for analyzing larger datasets and performing more advanced computations, such that the benefits of ML likely outweigh the errors that are inevitably produced in the ML models. Further clinical-research collaborations are required to fine-tune ML algorithms for different treatments and patient groups, and identify additional opportunities for ML applications to advance psychotherapy process and outcome. Clearly, there remains a need to consider ethics with regards to collecting, analyzing and sharing treatment data, as well as implementing ML based feedback tools into clinical practice. As ML algorithms continue to be refined and improved, it might be possible to help therapists to identify mental illnesses at an earlier stage when interventions may be more effective, and personalize treatments based on an individual's unique characteristics. Perhaps most importantly ML might enable therapists to focus on the relational aspects of psychotherapy that can only be achieved through the therapist–patient interactions.

Notes

[1] Prior to performing ML algorithms, Natural Language Processing (NLP; which is another type of AI) may be used to transform raw texts (i.e. clinical notes, session transcripts) to more useful labeled data. In essence, NLP methods take large collections of unstructured text as inputs, and with the use of computerized dictionaries, place specific words in psychologically meaningful categories (e.g. emotion words, reflecting or experiencing; e.g. Mergenthaler, 2008).

[2] Accuracy = the proportion of the total number of predictions that were correct; Precision = ratio of true positives and the total number of positives predicted by a model, also called exactness; Sensitivity = the proportion of positives correctly classified, also called recall or completeness; F_1 = harmonic mean of precision

and recall, where an F_1 score reaches its best value at 1 (perfect precision and recall); Specificity = The proportion of negatives correctly classified, which is the complete opposite of recall; ROC curve = receiver operating characteristic curve. A plot that shows the true positive rate against the false positive rate for various threshold values. Area under the curve (AUC) = the probability that the model ranks a random positive example more highly than a random negative example. Area representing the discriminative power of a test between 0.5 (no discrimination) and 1 (perfect discrimination; Dwyer et al., 2018; Sammut & Webb, 2017).

[3] Counts are not exclusive as several studies report on multiple study aims and approaches.

[4] In essence, NLP methods take large collections of unstructured text as inputs, and with the use of computerized dictionaries, place specific words in psychologically meaningful categories (e.g., emotion words, reflecting or experiencing; e.g., Mergenthaler, 2008).

[5] For example. if your dataset is already balanced (i.e. equal numbers of trials in each class and cross-validation fold, which is recommend) an evaluation metric of classification accuracy might suffice. However, when the dataset is imbalanced and the cost of misclassification of the minor class samples is very high, classification accuracy might give a false sense of achieving high accuracy. For example, if we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to do more tests (Rácz et al., 2019).

[6] See Chen and Wojcik (2016) for a practical guide to conducting big data research in psychology, covering data management, acquisition, processing, and analytics (including walkthrough tutorials on key supervised and unsupervised learning data mining methods).

Supplemental data

Supplemental data for this article can be accessed here https://doi.org/10.1080/10503307.2020.1808729.

ORCID

KATIE AAFJES-VAN DOORN http://orcid.org/0000-0003-2584-5897

References

Ali, M. R., Razavi, Z., Mamun, A. A., Langevin, R., Rawassizadeh, R., Schubert, L., & Hoque, M. E. (2019). A virtual conversational agent for teens with autism: Experimental results and design lessons. *ArXiv:1811.03046 [Cs]*. http://arxiv.org/abs/1811.03046.

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, *4*, 463–476.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, *26*(5), 816–827. https://doi.org/10.1037/a0029607

Atkins, D. C, Steyvers, M., Imel, Z. E, & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 49. https://doi.org/10.1186/1748-5908-9-49

Bacchetti, P. (2002). Peer review of statistics in medical research: The other problem. *BMJ: British Medical Journal*, *324*(7348), 1271–1273. https://doi.org/10.1136/bmj.324.7348.1271

Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S. C., Kong, D., Moody, A. R., & Tyrrell, P. N. (2019). Sample-size determination methodologies for machine learning in medical imaging research: A systematic review. *Canadian Association of Radiologists Journal*, *70*(4), 344–353. https://doi.org/10.1016/j.carj.2019.06.002

Ball, T. M., Stein, M. B., Ramsawh, H. J., Campbell-Sills, L., & Paulus, M. P. (2014). Single-subject anxiety treatment outcome prediction using functional neuroimaging. *Neuropsychopharmacology*, *39*(5), 1254–1261. https://doi.org/10.1038/npp.2013.328

Barrett, S. J., & Langdon, W. B. (2006). *Advances in the application of machine learning techniques in drug discovery, design and development*. 10th Online World Conference on Soft Computing in Industrial Applications (pp. 99–110). Springer.

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, *319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Berman, J. J. (2013). *Principles of big data: Preparing, sharing, and analyzing complex information*. Elsevier, Morgan Kaufmann.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology*, *ePub*. https://doi.org/10.1093/aje/kwz189

Big data. (2020). *Oxford Online Dictionary*. https://en.oxforddictionaries.com/definition/big_data.

Blamire, P. A. (1996). The influence of relative sample size in training artificial neural networks. *International Journal of Remote Sensing*, *17*(1), 223–230. https://doi.org/10.1080/01431169608949000

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., & Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, *13*(3), 158–174. https://doi.org/10.2196/jmir.1838

Caballero, Y., Bello, R., Taboada, A., Nowe, A., Garcia, M. M., & Casas, G. (2006, September). A new measure based in the rough set theory to estimate the training set quality. In *Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (pp. 133–140). IEEE. https://doi.org/10.1109/synasc.2006.6.

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, *63*(3), 343–350. https://doi.org/10.1037/cou0000111

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), 1–12. https://doi.org/10.1038/s41398-018-0355-8

Chaoua, I., Recupero, D. R., Consoli, S., & Harma, A. (2018). Detecting and tracking ongoing topics in psychotherapeutic conversations. 13.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/10.1037/met0000111

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology*, *14*(1), 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Cohen, Z. D., Kim, T. T., Van, H. L., Dekker, J. J. M., & Driessen, E. (2020). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*, *30*(2), 137–150. https://doi.org/10.1080/10503307.2018.1563312

Cohn, J. F., Cummins, N., Epps, J., Goecke, R., Joshi, J., & Scherer, S. (2018). Multimodal assessment of depression from behavioral signals. In *The handbook of multimodal-multisensor interfaces: Foundations, user modeling, and common modality combinations* (Vol. 2, pp. 375–417). Association for Computing Machinery. https://doi.org/10.1145/3107990.3108004

Crangle, C. E., Wang, R., Perreau-Guimaraes, M., Nguyen, M. U., Nguyen, D. T., & Suppes, P. (2019). Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the Stanford Suppes Brain Lab Psychotherapy Dataset. *ArXiv:1901.04110 [Cs, Eess]*. http://arxiv.org/abs/1901.04110.

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24. https://doi.org/10.1037/ccp0000476

Dibeklioğlu, H., Hammal, Z., & Cohn, J. F. (2017). Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE Journal of Biomedical and Health Informatics*, *22*(2), 525–536. https://doi.org/10.1109/JBHI.2017.2676878

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2019). Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*, *77*(1), 35–43.

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, *21*(5), e13216. https://doi.org/10.2196/13216

Flemotomos, N., Georgiou, P., & Narayanan, S. (2019). Language Aided Speaker Diarization Using Speaker Role Information. *ArXiv:1911.07994*. http://arxiv.org/abs/1911.07994.

Garcia, D., Archer, T., & Kostrzewa, R. M. (Eds.). (2019). *Personality and brain disorders: Associations and interventions*. Springer.

Gaut, G., Steyvers, M., Imel, Z. E., Atkins, D. C., & Smyth, P. (2017). Content coding of psychotherapy transcripts using labeled topic models. *IEEE Journal of Biomedical and Health Informatics*, *21*(2), 476–487.

Geraci, J., Wilansky, P., de Luca, V., Roy, A., Kennedy, J. L., & Strauss, J. (2017). Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evidence-Based Mental Health*, *20*(3), 83–87. https://doi.org/10.1136/eb-2017-102688

Gibson, J., Can, D., Xiao, B. S., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2016). A deep learning approach to modeling empathy in addiction counseling. *INTERSPEECH*, *2016*, 1447–1451. https://doi.org/10.21437/Interspeech.2016-554

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, *18*, 34–42. https://doi.org/10.1016/j.cobeha.2017.07.003

Gori, A., Lauro-Grotto, R., Giannini, M., & Schuldberg, D. (2010). Predicting treatment outcome by combining different assessment tools: Toward an integrative model of decision support in psychotherapy. *Journal of Psychotherapy Integration*, *20*(2), 251–269. https://doi.org/10.1037/a0019768

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports*, *21*(11), 116. https://doi.org/10.1007/s11920-019-1094-0

Hahn, T., Kircher, T., Straube, B., Wittchen, H.-U., Konrad, C., Ströhle, A., Wittmann, A., Pfleiderer, B., Reif, A., Arolt, V., & Lueken, U. (2015). Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. *JAMA Psychiatry*, *72*(1), 68–74. https://doi.org/10.1001/jamapsychiatry.2014.1741

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *American Journal of Roentgenology*, *212*(1), 38–43. https://doi.org/10.2214/AJR.18.20224

Hart, A., & Wyatt, J. (1990). Evaluating black-boxes as medical decision aids: Issues arising from a study of neural networks. *Medical Informatics*, *15*(3), 229–236. https://doi.org/10.3109/14639239009025270

Hasan, M. (2019). *Machine learning methods for the analysis of clinical conversation*. Wayne State University.

Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, *13*, 2813–2869.

Hirsch, T., Soma, C., Merced, K., Kuo, P., Dembe, A., Caperton, D. D., Atkins, D. C., & Imel, Z. E. (2018, June). *"It's hard to argue with a computer:" Investigating psychotherapists' attitudes towards automated evaluation*. Proceedings of the 2018 Designing Interactive Systems Conference (pp. 559–571). https://doi.org/10.1145/3196709.3196776

Hoogendoorn, M., Berger, T., Schulz, A., Stolz, T., & Szolovits, P. (2017). Predicting social anxiety treatment outcome based on therapeutic email conversations. *IEEE Journal of Biomedical and Health Informatics*, *21*(5), 1449–1459.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, *8*(7), 795–802. https://doi.org/10.1002/sim.4780080704

Idalski Carcone, A., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K., & Kotov, A. (2019). Developing machine learning models for behavioral coding. *Journal of Pediatric Psychology*, *44*(3), 289–299. https://doi.org/10.1093/jpepsy/jsy113

Imel, Z. E., Caperton, D. D., Tanana, M., & Atkins, D. C. (2017). Technology-enhanced human interaction in psychotherapy. *Journal of Counseling Psychology*, *64*(4), 385–393. https://doi.org/10.1037/cou0000213

Imel, Z. E., Pace, B. T., Soma, C. S., Tanana, M., Hirsch, T., Gibson, J., Georgiou, P., Narayanan, S., & Atkins, D. C. (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy*, *56*(2), 318–328. https://doi.org/10.1037/pst0000221

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, *52*(1), 19–30. https://doi.org/10.1037/a0036841

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR MHealth and UHealth*, *6*(11), e12106. https://doi.org/10.2196/12106

Intille, S. (2016). The precision medicine initiative and pervasive health research. *IEEE Pervasive Computing*, *15*(1), 88–91. https://doi.org/10.1109/MPRV.2016.2

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. https://doi.org/10.1007/978-1-4614-7138-7.

Jaroszewski, A. C., Morris, R. R., & Nock, M. K. (2019). Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *Journal of Consulting and Clinical Psychology*, *87*(4), 370–379. https://doi.org/10.1037/ccp0000389

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kim, S. (2008). Computer judgment of main color in a drawing for art psychotherapy assessment. *The Arts in Psychotherapy*, *35*(2), 140–150. https://doi.org/10.1016/j.aip.2008.01.002

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *The Lancet Psychiatry*, *3*(10), 935–946. https://doi.org/10.1016/S2215-0366(16)30171-7

Krause, K., Guertler, D., Moehring, A., Batra, A., Eck, S., Rumpf, H.-J., Bischof, G., Lucht, M., Freyer-Adam, J., Ulbricht, S., John, U., & Meyer, C. (2019). Feasibility and acceptability of an intervention providing computer-generated tailored feedback to target alcohol consumption and depressive symptoms in proactively recruited health care patients and reactively recruited media volunteers: Results of a pilot study. *European Addiction Research*, *25*(3), 119–131. https://doi.org/10.1159/000499040

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A.-K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier treatment navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438.

Månsson, K. N. T., Frick, A., Boraxbekk, C.-J., Marquand, A. F., Williams, S. C. R., Carlbring, P., Andersson, G., & Furmark, T. (2015). Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. *Translational Psychiatry*, *5*(3), e530. https://doi.org/10.1038/tp.2015.22

Mergenthaler, E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research*, *18*(2), 109–126. https://doi.org/10.1080/10503300701883741

Mikus, A., Hoogendoorn, M., Rocha, A., Gama, J., Ruwaard, J., & Riper, H. (2018). Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data. *Internet Interventions*, *12*, 105–110. https://doi.org/10.1016/j.invent.2017.10.001

Miner, A. S., Shah, N., Bullock, K. D., Arnow, B. A., Bailenson, J., & Hancock, J. (2019). Key considerations for incorporating conversational AI in psychotherapy. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00746

Modai, I., Israel, A., Mendel, S., Hines, E. L., & Weizman, R. (1996). Neural network based on adaptive resonance theory as compared to experts in suggesting treatment for schizophrenic and unipolar depressed in-patients. *Journal of Medical Systems*, *20*(6), 403–412.

Nasir, M., Baucom, B. R., Georgiou, P., & Narayanan, S. (2017). Predicting couple therapy outcomes based on speech acoustic features. *PLOS ONE*, *12*(9), e0185123. https://doi.org/10.1371/journal.pone.0185123

Nitti, M., Ciavolino, E., Salvatore, S., & Gennaro, A. (2010). Analyzing psychotherapy process as intersubjective sensemaking: An approach based on discourse analysis and neural networks. *Psychotherapy Research*, *20*(5), 546–563. https://doi.org/10.1080/10503301003641886

Pace, B., Tanana, M., Xiao, B., Dembe, A., Soma, C., Steyvers, M., & Imel, Z. E. (2016). What about the words? Natural language processing in psychotherapy. *Psychotherapy Bulletin*, *51*, 14–18.

Park, J., Kotzias, D., Kuo, P., Logan IV, R. L., Merced, K., Singh, S., Tanana, M., Karra Taniskidou, E., Lafata, J. E., Atkins, D. C., Tai-Seale, M., Imel, Z. E., & Smyth, P. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *Journal of the American Medical Informatics Association*, *26*(12), 1493–1504. https://doi.org/10.1093/jamia/ocz140

Rácz, A., Bajusz, D., & Héberger, K. (2019). Multi-level comparison of machine learning classifiers and their performance metrics. *Molecules*, *24*(15), 2811. https://doi.org/10.3390/molecules24152811

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(3), 252–264. https://doi.org/10.1109/34.75512

Redlich, R., Opel, N., Grotegerd, D., Dohm, K., Zaremba, D., Bürger, C., Münker, S., Mühlmann, L., Wahl, P., Heindel, W., Arolt, V., Alferink, J., Zwanzger, P., Zavorotnyy, M., Kugel, H., & Dannlowski, U. (2016). Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. *JAMA Psychiatry*, *73*(6), 557–564. https://doi.org/10.1001/jamapsychiatry.2016.0316

Reggente, N., Moody, T. D., Morfini, F., Sheen, C., Rissman, J., O'Neill, J., & Feusner, J. D. (2018). Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive–compulsive disorder. *Proceedings of the National Academy of Sciences*, *115*(9), 2222–2227. https://doi.org/10.1073/pnas.1716686115

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv:1602.04938 [Cs, Stat]*. http://arxiv.org/abs/1602.04938.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell, F. E., Jr, Moons, K. G., & Collins, G. S. (2019a). Minimum sample size for developing a multivariable prediction model: Part I - continuous outcomes. *Statistics in Medicine*, *38*(7), 1262–1275. https://doi.org/10.1002/sim.7993

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell, F. E., Jr, Moons, K. G., & Collins, G. S. (2019b). Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. *Statistics in Medicine*, *38*(7), 1276–1296. https://doi.org/10.1002/sim.7992

Rubel, J. A., Zilcha-Mano, S., Giesemann, J., Prinz, J., & Lutz, W. (2019). Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration. *Psychotherapy Research*, *30*(3), 300–309. https://doi.org/10.1080/10503307.2019.1597994

Sahiner, B., Chan, H. P., & Hadjiiski, L. (2008). Classifier performance estimation under the constraint of a finite sample size: Resampling schemes applied to neural network classifiers. *Neural Networks*, *21*(2-3), 476–483. https://doi.org/10.1016/j.neunet.2007.12.012

Sammut, C., & Webb, G. I. (2017). *Encyclopedia of machine learning and data mining*. Springer. https://doi.org/10.1007/978-0-387-30164-8.

Schmitgen, M. M., Niedtfeld, I., Schmitt, R., Mancke, F., Winter, D., Schmahl, C., & Herpertz, S. C. (2019). Individualized treatment response prediction of dialectical behavior therapy for borderline personality disorder using multimodal magnetic resonance imaging. *Brain and Behavior*, *9*(9), e01384. https://doi.org/10.1002/brb3.v9.9

Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2020). Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research*, 1–19. https://doi.org/10.1080/10503307.2020.1769219

Seuchter, S. A., Eisenacher, M., Riesbeck, M., Gaebel, W., & Köpcke, W. (2004). Methods for predictor analysis of repeated measurements: Application to psychiatric data. *Methods of Information in Medicine*, *43*(2), 184–191. https://doi.org/10.1055/s-0038-1633857

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Shiner, B., D'Avolio, L. W., Nguyen, T. M., Zayed, M. H., Young-Xu, Y., Desai, R. A., Schnurr, P. P., Fiore, L. D., & Watts, B. V. (2013). Measuring use of evidence based psychotherapy for posttraumatic stress disorder. *Administration and Policy in Mental Health and Mental Health Services Research*, *40*(4), 311–318. https://doi.org/10.1007/s10488-012-0421-0

Silberschatz, G. (2017). Improving the yield of psychotherapy research. *Psychotherapy Research*, *27*(1), 1–13. https://doi.org/10.1080/10503307.2015.1076202

Singla, K., Chen, Z., Flemotomos, N., Gibson, J., Can, D., Atkins, D., & Narayanan, S. (2018). Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. *Interspeech*, *2018*, 3413–3417. https://doi.org/10.21437/Interspeech.2018-2551

Symons, M., Feeney, G. F. X., Gallagher, M. R., Young, R. McD., & Connor, J. P. (2019). Machine learning vs addiction therapists: A pilot study predicting alcohol dependence treatment outcome from patient data in behavior therapy with adjunctive medication. *Journal of Substance Abuse Treatment*, *99*, 156–162.

Tai, A. M. Y., Albuquerque, A., Carmona, N. E., Subramanieapillai, M., Cha, D. S., Sheko, M., Lee, Y., Mansur, R., & McIntyre, R. S. (2019). Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artificial Intelligence in Medicine*, *99*, 101704. https://doi.org/10.1016/j.artmed.2019.101704

Tajika, A., Ogawa, Y., Takeshima, N., Hayasaka, Y., & Furukawa, T. A. (2015). Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *The British Journal of Psychiatry*, *207*(4), 357–362. https://doi.org/10.1192/bjp.bp.113.143701

Tanana, M., Hallgren, K., Imel, Z., Atkins, D., Smyth, P., & Srikumar, V. (2015). Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 71-79).

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, *65*, 43–50. https://doi.org/10.1016/j.jsat.2016.01.006

Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *Journal of Medical Internet Research*, *21*(7), e12529. https://doi.org/10.2196/12529

Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: a Multidisciplinary Journal*, *20*(4), 640–657. https://doi.org/10.1080/10705511.2013.824781

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, *2*(1), 1–14. https://doi.org/10.1002/jrsm.27

Tolmeijer, E., Kumari, V., Peters, E., Williams, S. C. R., & Mason, L. (2018). Using fMRI and machine learning to predict symptom improvement following cognitive behavioural therapy for psychosis. *Neuro Image Clinical*, *20*, 1053–1061.

Tseng, S.-Y., Baucom, B., & Georgiou, P. (2017). Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings. *Interspeech*, *2017*, 3291–3295. https://doi.org/10.21437/Interspeech.2017-1621

Villmann, T., Liebers, C., Bergmann, B., Gumz, A., & Geyer, M. (2008). Investigation of psycho-physiological interactions between patient and therapist during a psychodynamic therapy and their relation to speech using in terms of entropy analysis using a neural network approach. *New Ideas in Psychology*, *26*(2), 309–325. https://doi.org/10.1016/j.newideapsych.2007.07.010

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

Warner, H. R., Toronto, A. F., Veasey, L. G., & Stephenson, R. (1961). A mathematical approach to medical diagnosis: Application to congenital heart disease. *JAMA*, *177*(3), 177–183. https://doi.org/10.1001/jama.1961.03040290005002

Watts, B. V., Shiner, B., Zubkoff, L., Carpenter-Song, E., Ronconi, J. M., & Coldwell, C. M. (2014). Implementation of evidence-based psychotherapies for posttraumatic stress disorder in VA specialty clinics. *Psychiatric Services*, *65*(5), 648–653. https://doi.org/10.1176/appi.ps.201300176

Weusthoff, S., Gaut, G., Steyvers, M., Atkins, D. C., Hahlweg, K., Hogan, J., Zimmermann, T., Fischer, M. S., Baucom, D. H., Georgiou, P., Narayanan, S., & Baucom, B. R. (2018). The language of interpersonal interaction: An interdisciplinary approach to assessing and processing vocal and speech data. *The European Journal of Counselling Psychology*, *7*(1), 69–85. https://doi.org/10.5964/ejcop.v7i1.82

Xiao, B., Georgiou, P., Baucom, B., & Narayanan, S. S. (2015). Head motion modeling for human behavior analysis in dyadic interaction. *IEEE Transactions on Multimedia*, *17*(7), 1107–1119. https://doi.org/10.1109/TMM.2015.2432671

Xiao, B., Huang, C., Imel, Z. E., Atkins, D.d C., Georgiou, P., & Narayanan, S. S. (2016). A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science*, *2*, e59. https://doi.org/10.7717/peerj-cs.59

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., Narayanan, S.S., & Sakakibara, M. (2015). "Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLOS ONE*, *10*(12), e0143055. https://doi.org/10.1371/journal.pone.0143055

Xing, Z., Zhao, X., & Miao, C. (2017). *Identifying cognitive distortion by convolutional neural network based text classification*. https://dr.ntu.edu.sg//handle/10356/89482.

Yang, Z., Wu, J., Xu, L., Deng, Z., Tang, Y., Gao, J., Hu, Y., Zhang, Y., Qin, S., Li, C., & Wang, J. (2019). Individualized psychiatric imaging based on inter-subject neural synchronization in movie watching. *NeuroImage*, *216*, 116227. https://doi.org/10.1016/j.neuroimage.2019.116227

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zilcha-Mano, S., Errázuriz, P., Yaffe-Herbst, L., German, R. E., & DeRubeis, R. J. (2019). Are there any robust predictors of "sudden gainers," and how is sustained improvement in treatment outcome achieved following a gain? *Journal of Consulting and Clinical Psychology*, *87*(6), 491–500. https://doi.org/10.1037/ccp0000401

Zilcha-Mano, S., Roose, S. P., Brown, P. J., & Rutherford, B. R. (2018). A machine learning approach to identifying placebo responders in late-life depression trials. *The American Journal of Geriatric Psychiatry*, *26*(6), 669–677. https://doi.org/10.1016/j.jagp.2018.01.001