Artificial Intelligence, Machine Learning and Mental Health

Jaime Delgadillo[1] and Dana Atzil-Slonim[2]

1. Clinical and Applied Psychology Unit, Department of Psychology, University of Sheffield, United Kingdom [*correspondence*: j.delgadillo@sheffield.ac.uk]
2. Department of Psychology, Bar-Ilan University, Israel

Abstract

Computers are capable of learning how to solve complex problems. The emergence of machine learning (ML) represents a major advance for the field of mental health. ML algorithms can be trained to recognize subgroups of people with similar symptoms (diagnosis), to estimate the probability of recovery from these symptoms (prognosis), to make a judgement about the best treatment option for a patient (treatment selection), and even to provide feedback and guidance to therapists by learning from recordings of effective therapist-patient interactions (process feedback). This article offers an introduction to ML and the emerging field of precision mental health care.

Keywords

Artificial intelligence; AI; Machine learning; Statistical learning; Deep learning; Data mining; Data science; Clinical prediction models; Natural language processing; Precision medicine; Precision mental health care

Objectives

- To explain machine learning and related concepts
- To explain the relationship between machine learning and artificial intelligence
- To discuss the relevance of the above concepts for mental health
- To provide examples of how artificial intelligence is being used to inform and deliver mental health care

Introduction: Explanatory and algorithmic models

Modern health care has advanced tremendously through the application of the scientific method, characterized by the generation of hypotheses from clinical observations and their subsequent testing using experimental methods such as controlled trials. Empirical tests help to refine clinical practice by confirming or disconfirming prior assumptions about the effects of interventions and their mechanisms of action. This hypothesis-testing approach is supported by the use of conventional statistical methods, which we refer to as *explanatory models* (Breiman, 2001). These models aim to explain relationships between inputs (i.e., independent variables) and outputs (i.e., dependent variable), and they assume that the outputs are produced by some process or mechanism (i.e., an underlying biological and/or psychological phenomenon). In recent decades, computers have become much more efficient at collecting and processing large volumes of data on human activity, inspiring statistical theorists and computer scientists to advance new methods to interrogate complex datasets. These developments have given rise to a new generation of *algorithmic models* (Breiman, 2001). These models aim to discover stable patterns of relationships between inputs in order to generate a set of outputs. Typically, these outputs are in the form of a prediction (e.g., probability of recovery from a mental health problem), a classification (e.g., diagnostic category) or a response (e.g., a text-based answer to a question). This 'data mining' process makes no assumptions about underlying mechanisms and can indeed be 'hypothesis free' and theoretically agnostic: it will use any inputs that are available in order to generate an output. Algorithmic models prioritize the specification of formal (i.e., mathematical, logical) rules that will ultimately help to maximize accuracy and to minimize error when producing an output. Sometimes these formal rules can be highly complex, making the model a 'black box'. As such, algorithmic models do not prioritize the 'explanation' of relationships between variables or underlying mechanisms, they prioritize the accuracy and utility of the output. *Machine learning* refers to a data mining process used to train computerized algorithms to solve prediction and classification problems in order to generate a useful output. Data sources used to develop a ML model are referred to as 'training' data. Once a model is trained, it can be 'fed' new data samples in order to produce an output. A trained algorithm can then be instantiated into software to automate this input-output process, which constitutes the basis for artificial intelligence (AI) technologies that help to solve problems or perform tasks. This article offers a primer on machine learning (ML) and related concepts in the context of mental health care.

Broad types of machine learning models

Contemporary textbooks in the fields of statistics and data science distinguish between two types of statistical learning: *supervised* and *unsupervised* learning (Hastie, Friedman, & Tibshirani, 2009). In

supervised ML, the training inputs include data records that have both 'features' and 'labels or values'. Features refer to variables that characterize a sample (e.g., demographics, symptoms, biomarkers, neurocognitive tests and personality traits of a clinical sample). Labels and values represent the outputs of interest. For example, using the features listed above, a model could be trained to output a diagnosis (a categorical label) or to predict the expected level of psychological distress after treatment (a value on a continuous scale of distress). In order to achieve this, the training sample should contain labelled data, where the diagnosis and post-treatment distress severity are available. These labelled samples serve as examples that the model uses to learn about the pattern of features that are usually associated with a specific label or value. A related concept is that of 'discriminant ML models', whose primary task is to recognize labels that have been learned through a training process.

In unsupervised ML, the training inputs include features but not labels, since the output is not known or specified a priori. The aim is to discover naturally occurring clusters or subgroups in the data. For example, using the features listed above, patients could be classified into latent clusters of cases with highly similar clinical-demographic profiles. Thus, the output of an unsupervised ML model is a predicted classification which is generated (i.e., discovered) by the data mining process. A related concept is that of 'generative ML models', which are capable of learning underlying probabilistic rules that model the distribution of records in the data space, and which are capable of generating (i.e., simulating) new data using those rules. Moreover, 'semisupervised' ML models combine insights from datasets where some cases are labelled and others are not (Zhu & Goldberg, 2009); for instance, where diagnostic labels are only available for some cases but the same features are available for all cases.

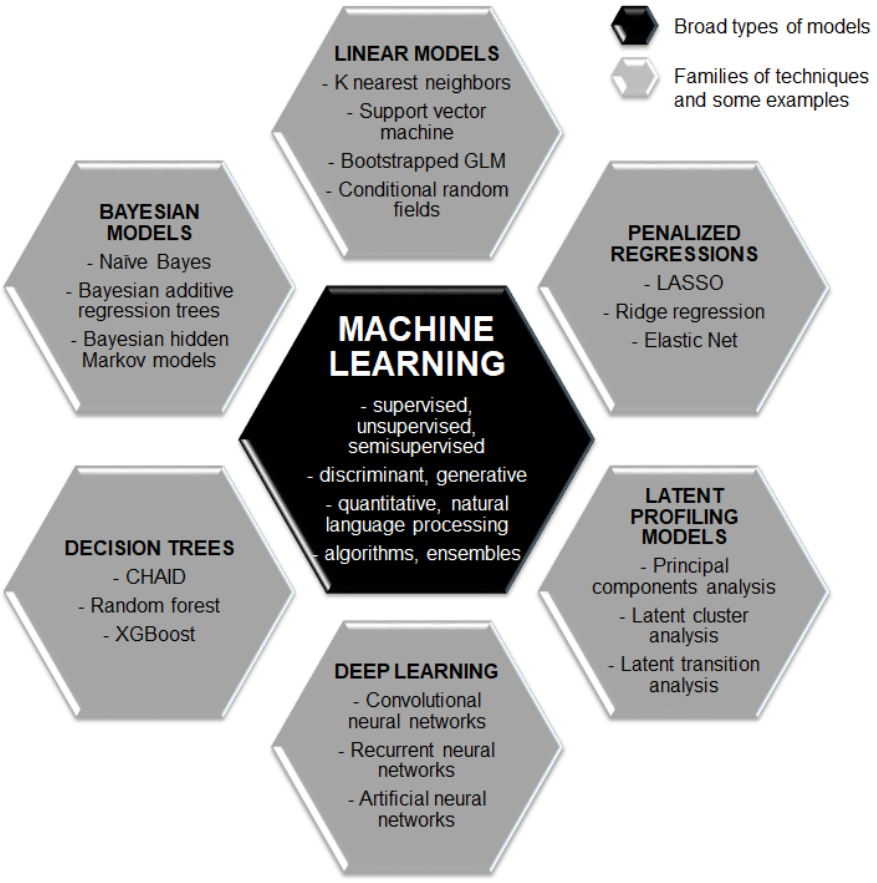
Many applications of ML make use of structured quantitative data, where the features and labels are coded into categorical, ordinal or continuous variables. Furthermore, the maturation of subfields of computer science such as natural language processing (NLP), signal processing (SP), and computer vision (CV), allow the analysis of massive amounts of complex data and can lead to dramatic progress in studying mental health problems. For example, audio recordings and text-based transcripts from many psychotherapy sessions can be used to identify therapist (verbal) behaviors that predict post-session symptomatic improvements using a supervised ML model. 'Sentiment analysis' of patient verbal behaviors could be used to train a supervised ML model to identify the label that best represents their emotional state. Or an unsupervised ML model like 'topic modeling' can be used to automatically identify regularly occurring classes of therapist behaviors throughout therapy.

Once a ML model is trained and its data processing rules are 'locked' (no longer continue to be trained), it is referred to as an 'algorithm' that is capable of generating an output when it receives

incoming data. An 'ensemble' refers to a ML model that consists of multiple algorithms that together produce an output. For example, the final predicted value could be the mean value across all algorithms ('model averaging'), or the final predicted classification can be the category that has the majority vote from all algorithms ('model voting'). There are three common ensembling techniques:

- *Bagging*: 'Bootstrap aggregation' involves training numerous algorithms from the same ML family in multiple bootstrapped datasets generated from an original dataset.
- *Boosting*: The serial training of models to derive strong learners from weak learners. Numerous models from the same ML family are produced, some of which are more or less accurate.
- *Stacking*: This involves combining the results of heterogeneous algorithms from different ML families.

Figure 1. Machine learning: types of models and families of techniques



Families of machine learning techniques

There are many ML modeling techniques, and variations thereof, which makes it challenging to provide a comprehensive taxonomy. Nevertheless, these are commonly grouped according to broad families of statistical techniques. Figure 1 provides a summary of broad types of ML models and families of ML techniques with some common examples.

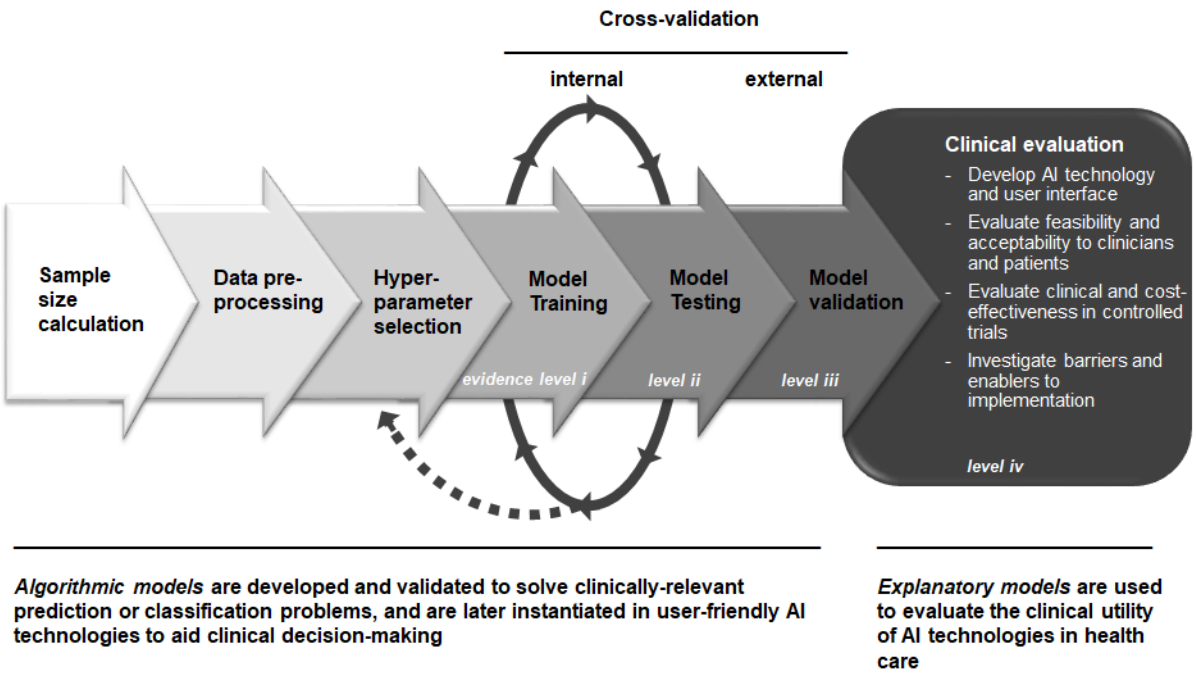
For example, Chi-square Automatic Interaction Detector (CHAID), Random Forest, and Extreme Gradient Boosting (XGBoost) are different techniques that belong to the wider family of Decision Trees. Decision trees are statistical models that help to discover interactions between variables, thus identifying subgroups of cases with similar features, who tend to have similar values in the dependent variable of interest. Another feature of decision trees is that they can model nonlinear associations between independent and dependent variables.

Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression, and Elastic Net are models that use a technique called regularization and which belong to the wider family of Penalized Regression Models. Regularization is a process that can help to perform variable selection (e.g., identify the most statistically reliable predictors) and weight-setting (e.g., attribute different weights to predictors, according to their predictive value). This is achieved by penalizing (shrinking) regression coefficients in order to arrive at an optimal model that minimizes overfitting to the training dataset and maximizes prediction accuracy.

Latent cluster analysis and latent transition analysis are unsupervised machine learning techniques that belong to the wider family of Latent Profiling Models. These methods enable the data-driven discovery of latent classes, factors or phenotypes. This is achieved by identifying the subgroups of cases that have highly similar values across a set of variables, and the parsing of subgroups can be performed empirically (e.g., any number of mutually exclusive classes can emerge from the data) or can be constrained based on prior evidence or theory (e.g., a specific number or a maximum number of classes could be extracted).

These examples illustrate how different ML techniques are geared to achieve specific goals (e.g., to model interactions, model nonlinear relationships, select useful predictors from a list of candidate variables, reduce risk of overfitting, discover latent classes, etc.). Hence, the selection of an appropriate ML technique should be made based on the specific goals of the analysis and the features and constraints of the available data (e.g., sample size, number of predictors, expectations about the relevance of interactions and nonlinear relationships, etc.).

Figure 2. Steps of a development pipeline for the clinical implementation of AI



Machine learning pipelines

Training a ML model involves several decisions and steps along a 'pipeline' of model development and validation, represented in Figure 2. High-level objectives are to train a model using data inputs and to evaluate its performance. Other sub-tasks involve the collection of adequately sized training and test datasets, the preparation of data for analysis, the selection of hyperparameters, cross-validation procedures, model evaluation and clinical field-testing. In this section, we provide brief descriptions of these sub-tasks of a ML pipeline.

Sample size calculation

Like any other statistical model, the usefulness of a ML algorithm depends on the quality (i.e., are the features relevant to the prediction task and measured reliably?) and quantity (i.e., is the dataset adequately powered?) of training samples. The sample size requirements will vary according to parameters such as the expected magnitude of association between inputs-output (i.e., explained variance or effect size), the number of features in the dataset, the expected magnitude of prediction shrinkage (i.e., reduction of accuracy in a test sample) and the particular ML technique(s) that will be applied. We refer the reader to generic sample size calculation guidelines for the development of multivariable prediction models (Riley et al., 2019a, 2019b) and for their external validation (Archer et al., 2020). There are also model-specific guidelines for techniques such as neural networks (Blamire,

1996; Sahiner et al., 2008), latent profile analysis (Tein et al., 2013), decision trees (Morgan et al., 2003) and others.

Data pre-processing and feature engineering

Pre-processing refers to data preparation tasks that are required prior to initiating model training. In the context of structured quantitative data, there are several considerations and tasks that may be required such as those listed below.

- Handling of missing data (e.g., applying imputation of data missing at random, only analyzing cases with complete data, or treating missing data as an informative category).
- Reduction of categorical data (e.g., collapsing/merging categories with small samples or where categories are not significantly different) or one-hot-encoding (e.g., treating each category as a separate binary variable [0 = absent; 1 = present], such as each instance of a word in a dataset that transforms therapy session transcript data into quantitative variables).
- In the context of generalized linear models, training data with highly skewed variables and extreme outliers may adversely influence the generalizability of a prediction model and some transformations may be required (e.g., normalization of variables, Winsorization of outliers by replacing the extreme values with the maximum or minimum value at a relevant threshold).
- Handling of class imbalance (e.g., retaining the original base rates of target categories, or using resampling techniques to correct for class imbalance).
- Handling naturalistic data where treatments were not assigned randomly (e.g., retaining the original treatment samples, or applying case-control matching techniques to balance covariates across treatments).

In the context of text-based ML analysis, some of the following considerations and pre-processing tasks are relevant.

- Tokenizing – dividing the text into unique units, called tokens. The term 'token' refers to the total number of words in a corpus regardless of how often they are repeated. The term 'type' refers to the number of distinct words in the text.
- Stop words removal – getting rid of common language articles, pronouns and prepositions such as 'and', 'the' or 'to' in English. Very common words that appear to provide little or no value to the NLP objective are filtered and excluded from the text to be processed, hence removing widespread and frequent terms that are not informative about the corresponding text.
- Stemming and lemmatization – both have the objective of reducing a word to its base or root form. The 'Stemming' algorithm identifies the common root form of a word by removing or

replacing word affixes (such as “ing”, “s”, “es”). This procedure is fast and simple, but sometimes this can result in a word that is not part of the language (e.g., “studies” is stemmed as “studi”). ‘Lemmatization’ ensures the ‘lemma’ will be a word that exists in the language, and considers the full vocabulary of a language to apply a morphological analysis to words (e.g., ”studies” is lemmatized as ”study”, ”was” as ”be”, and ”better” as ”good”).

- Parts of speech tagging – tagging words to nouns, verbs, etc.
- Text representation – since ML models are only capable of processing numerical values, the tokens in a sentence are replaced by numbers. There are various techniques to convert words into numbers. One of the most basic techniques used to represent textual data is ‘one-hot-encoding’. In this technique, a vector is created in the size of the total number of unique words. The value of vectors is assigned such that the value of each word belonging to its index is 1 and the others are 0. For example, here is a one hot vector representation of the sentence: “I feel very anxious”:

	I	feel	very	anxious
I	1	0	0	0
feel	0	1	0	0
very	0	0	1	0
anxious	0	0	0	1

As illustrated in this example, every word has its own value in a vector. This technique is easy to implement, but it does not take the semantic meaning and context into account. More advanced text representation techniques, like ‘word embedding’ take the semantic context into account and give words with similar meaning or influence in a sentence similar values. With word embedding, each word is represented by a dense vector of fixed size (generally range from 50 to 300), with values corresponding to a set of features representing different aspects of a word’s semantic meaning (e.g., royalty, gender, plural, etc.). These features are obtained by random initialization, and are updated by the model during training. As can be seen in the following example, words with similar meaning, such as king and queen, would end up being closer to one another. In addition, the semantic relationship between different embeddings can be illustrated by the similar distance between “queen” and “king” to the distance between “woman” and “man”.

	Living being	human	gender	royalty	verb	plural
kitten	0.5	-0.1	0.2	-0.6	-0.5	-0.1
houses	-0.8	-0.5	0.1	-0.9	0.3	0.8
man	0.6	0.8	0.9	-0.1	-0.9	-0.7
woman	0.7	0.9	-0.7	0.1	-0.5	-0.4
king	0.5	0.7	0.8	0.9	-0.7	-0.6
queen	0.8	0.8	-0.9	0.8	-0.5	-0.9

In the context of audio and voice-based ML analysis, some of the following considerations and pre-processing tasks are relevant:

- Noise reduction – is a pre-processing procedure used to delineate the spoken signal from other sounds that may populate the recordings. The process usually utilizes the low-pass filters on the frequency domain based on the knowledge that human speech pitch is lower than 500Hz.
- Voice activity detection – distinguishes between moments of speech from moments of silence (see Li et al, 2016).
- Diarization – distinguishes between segments of the speakers' turns, and segments of overlapped speech (see Laufer-Goldshtein et al., 2018).
- Feature extraction – for each speech turn, there can be several speech features that can be extracted. Examples of common features used in speech analysis are f0 (the fundamental frequency or main pitch) the intensity (the volume of the speaker) as well as speech rate and speech quality features such as HF500 (the ratio between the main frequency of the speech to higher frequency relating to speech quality (see Bone et al., 2014).

Hyperparameter selection and tuning

Unlike conventional statistical tests, where there is a standard way of computing a result (e.g., performing a comparison of means using t-tests), ML techniques require the analyst to specify a number of relevant 'rules' that will guide the training process. Hyperparameters can be thought of as these 'rules' or 'settings' that determine how a model will learn from data. ML techniques such as those listed in Figure 1 have their particular hyperparameters. For instance, when training an ensemble of decision trees, the analyst has to decide: how many decision trees will be generated, the maximum tree depth (levels), the minimum parent-to-child node size (sample size), p-value

significance level to determine splitting and merging, whether or not to apply Bonferroni correction for multiple testing, etc.

Hyperparameters are usually selected a priori. However, the analyst may decide to manually 'tune' the hyperparameters to learn how different settings influence model performance. If this manual tuning is performed unsystematically, it can result in overfitting and poor model generalization. For these reasons, automated 'grid searching' methods can be applied to search for the optimal configuration of hyperparameters that maximizes model performance.

Cross-validation

Cross-validation (CV) is a central concept in the field of ML. It involves using some samples to train a model and other samples to evaluate its performance. This can be achieved by using different data sources for the training-evaluation steps, or by partitioning a large dataset into separate subsets using split-half (50% training, 50% test) or imbalanced-split (i.e., 70:30 train-test, or 50:25:25 train-test-validation) sampling techniques. Ensuring that training and evaluation samples are completely separate is essential to minimize 'information leakage', which can lead to overoptimistic evaluations of the performance of a ML model (Kocak et al., 2021). Hence, it is advisable to use different samples (datasets or partitions) for training, testing and validation purposes.

Internal CV refers to a process where a model is trained and evaluated using a single data source, which is known to be bias-prone due to information leakage and overfitting. Techniques to minimize these biases in internal CV include k-fold CV (split into k pieces, typically 5 or 10), leave-one-out (a single test case is held-out of the training sample in iterative loops), or bootstrapping methods (Efron & Tibshirani, 1997; Rodriguez, Perez, & Lozano, 2009). These internal CV techniques usually involve multiple training-test iterations (e.g., 10 iterations in 10-fold CV), represented by a clockwise loop in Figure 1. Internal CV loops are often used to perform sub-tasks that optimize model performance, such as hyperparameter tuning and variable selection.

External CV or 'model validation' refers to a process where a trained model's performance is evaluated in a completely independent or 'hold-out' sample. This method offers a more robust test of generalizability, as it represents the extent to which a model could be used in clinical care to make predictions or classifications using data from new patients.

Evaluation and transparent reporting of ML model performance

Most of the steps and tasks outlined above are geared towards maximizing prediction/classification accuracy and minimizing error. Specific indices of accuracy and error are used to train (i.e., during internal CV) and evaluate a model, depending on the type of output (for a detailed discussion, see Handelman et al., 2019; Kocak et al., 2021). For example, binary classifiers are usually evaluated using the Area Under the Curve (AUC) statistic, which is derived from a Receiver Operating Characteristic

(ROC) curve analysis. ROC curve analysis is commonly used to evaluate the performance of medical diagnosis and screening tests, by examining the full space of possible cut-offs in a continuous measure (e.g., an index of disease severity) with their corresponding sensitivity and specificity values. The AUC is a summary statistical indicator of the test's accuracy, where a value of 0.50 indicates chance-level performance and values closer to 1.0 are indicative of better accuracy. In addition to the AUC, a transparent evaluation of performance accuracy should also report the model's corresponding positive and negative predictive values and confusion matrix (which can help to derive other metrics, such as balanced accuracy). Models that predict continuous values are usually evaluated by reporting the absolute correlation between predicted-observed values, R squared, mean squared error and root mean squared error. In addition, calibration plots enable a visual inspection of the correspondence between predicted and observed values in a test or validation sample across the full range of predicted values. As a general rule, authors should report indices of accuracy, error and calibration plots within the training and test/validation samples to aid interpretability. Reporting indices of performance accuracy in both the training and test/validation samples is helpful to understand the extent to which prediction shrinkage occurs when a trained model is applied in a different sample. For instance, two different machine learning models could obtain the same performance accuracy (e.g., AUC = 0.65) in a hold-out sample, but their performance in the training sample could differ substantially (Model 1 AUC = 0.68; Model 2 AUC = 0.89). In this example, we would have greater confidence in Model 1, which shows lower prediction shrinkage (AUC shrinkage = 0.03 vs. 0.24) and therefore higher stability when applied out of sample. Model 2 is likely to be overfitting to the training sample, and might not generalize well to a new population in clinical practice.

The strength and credibility of evidence for the clinical utility of ML-based AI technologies ranges on a continuum, as illustrated in Figure 2. *Level i* evidence is the least stringent, and it simply involves testing the performance (prediction/classification accuracy) of a ML model within the sample used to train it. *Level ii* evidence improves upon the prior level by including an internal CV procedure, where performance is assessed in out-of-sample cases, partitions or bootstraps (depending on the internal CV approach used). *Level iii* evidence uses a statistically independent hold-out sample for model validation, and could be based on a training-validation approach or a training-testing-validation approach (if, for example, hyperparameters are tuned using internal CV using the training-testing partitions).[*] *Level iii* is the highest level of evidence within the *algorithmic modeling* paradigm, which

[*] *Note:* We acknowledge that when three datasets are used (A-B-C), and where the intermediate (B) set is used to perform tasks such as variable selection and hyperparameter tuning, some authors refer to these as training-validation-test sets. We argue that it is more logical to refer to these as training-test-validation sets, since the B set is serving the preliminary function of *testing* how the performance is impacted by different settings or variables. The expression *validation* denotes a more definitive evaluation of the statistical and/or clinical value of a model or technology (i.e., *level iii* evidence), and it aligns with the more commonly used concept of clinical validation (which provides *level iv* evidence).

could justify a transition towards clinical field-testing using conventional healthcare evaluation methods and *explanatory models* of statistical analysis (i.e., hypothesis-testing). The final stage in the pipeline towards implementation involves the evaluation of AI technologies in health care (or other intended user) populations, which yields the most stringent level of evidence (*Level iv*). Such evidence could include data on the feasibility, acceptability, clinical and cost-effectiveness of AI-driven interventions, supporting the case for implementation in routine care.

Given the complexities involved in developing ML models, there are several sources of bias that can occur at each step of the modeling pipeline. We refer readers to articles by Delgadillo (2021) and Kocak et al. (2021) which provide detailed tables of common sources of bias, which can be useful for developers and peer reviewers of ML studies. In addition, there are published guidelines for the transparent reporting of clinical prediction model development studies (e.g., Collins et al., 2015) and the appraisal of methodological quality for systematic reviews of these studies (Wolff et al., 2019). Furthermore, the development of AI-specific reporting guidelines is currently underway (Collins et al., 2021; Sounderajah et al., 2021).

The relevance of machine learning for mental health

Many important decisions in the field of mental health are made based on theory or clinical experience and intuition. Knowledge and interpretation of theory, as well as clinical judgment, vary considerably between clinicians and within clinicians (i.e., over time). For these reasons, it is not surprising that there is considerable variability in the accuracy of diagnostic and prognostic judgments between clinicians, whereas statistical models have been found to be at least as accurate and often superior to expert clinicians (see meta-analysis by Ægisdóttir et al., 2006). Similarly, numerous studies in the field of psychotherapy have found that clinical outcomes vary considerably between therapists (see seminal meta-analysis by Baldwin and Imel, 2013). It is not surprising to see variability in treatment response in situations where therapists may follow an eclectic or integrative approach, since some therapists may have a better intuition for which interventions or interactions may be most suitable for each patient. Yet, outcome variability between therapists persists even in situations where therapists are delivering the same intervention to a homogeneous clinical group (e.g., patients with major depressive disorder), such as in controlled trials of empirically supported psychotherapies. Overall, it is evident that there is considerable room for improvement in clinical decisions concerning diagnosis, prognosis, treatment selection and delivery/adaptation for patients with mental health problems.

Recent trends such as the collection of 'big data' in mental health services and the growing use of ML techniques have led researchers in the field to call for the development of a data-informed

approach to improve diagnosis and treatment – signaling the emergence of a new field of *precision mental health care* (Bzdok & Meyer-Lindenberg, 2018; Delgadillo & Lutz, 2020; DeRubeis, 2019; Fernandes et al., 2017; Kessler & Luedtke, 2021; Lutz et al., 2022). A comprehensive review of this literature is beyond the scope of the present article, so we refer readers to systematic and narrative reviews in the fields of neuroimaging, psychiatry and psychotherapy (Aafjes-van Doorn et al., 2021; Chekroud et al., 2021; Dwyer et al., 2018; Meehan et al., 2022; Salazar de Pablo et al., 2021; Shatte et al., 2019; Walter et al., 2019). In the next sections, we provide some selected examples to illustrate the types of clinical problems and applications that have been approached using ML techniques.

Diagnostic models

Initiatives such as the Research Domain Criteria (RDoC) have propelled the collection of multi-domain information including genetic, psychosocial, symptomatic, neurocognitive, biometric and passive sensing data (e.g., activity data from smartphones) to refine our understanding of mental health problems (Insel et al., 2010; Torous et al., 2017). The analysis of such high-dimensional datasets using ML techniques could enable the data-driven classification of stable clinical phenotypes. For example, Orru et al. (2012) reviewed over 40 studies that applied a Support Vector Machine model to identify imaging biomarkers of neurological and psychiatric disorders, generally indicating clinically acceptable performance across studies reporting heterogeneous indices of classification accuracy. Whelan et al. (2014) analyzed multi-domain data using Elastic Net to characterize the neuropsychosocial profiles of adolescents with alcohol use disorders. In another example, the 24th Machine Learning for Signal Processing (MLSP) competition required participants to automatically diagnose schizophrenia using a labelled dataset including high-dimensional features derived from MRI scans; the results indicated that ML techniques could achieve diagnostic classification with a maximum accuracy of AUC = 0.88 (Silva et al., 2014). More recently, two separate studies used the same unsupervised ML technique (Hidden Markov Models) to identify different diagnostic subtypes of depression (Catarino et al., 2022; Simmonds-Buckley et al., 2021). They yielded convergent findings that a specific subgroup of patients with a 'somatic depression' subtype tended to respond less well to Cognitive Behavioral Therapy (CBT).

Prognostic models

One of the earliest applications of ML techniques in the mental health literature trained a K-Nearest Neighbors (KNN) model to predict psychological treatment outcomes (Lutz et al., 2005). The KNN technique identifies cases with highly similar features (i.e., *neighbors*) to each individual case represented in a dataset, and then predicts an outcome of interest for each individual using data from their nearest neighbors. Since this seminal study, there has been an upsurge in the development of models that aim to predict future events or health states using pre-treatment data. For example,

Sajjadian et al. (2021) reviewed 54 studies that developed models to predict antidepressant treatment response for patients with major depressive disorder. The studies applied various ML techniques such as random forest,
extreme gradient boosting, LASSO regularization, elastic net, naïve Bayes, support vector machine, and others. The reviewed studies reported promising indices of prediction accuracy, although some studies had relatively small samples and limited evidence of external cross-validation (internal AUC = 0.71-0.86; external AUC = 0.70-0.79). In the field of psychotherapy, pre-treatment data have been analyzed using supervised ML techniques to predict treatment response (e.g., Flygare et al., 2020; Hilbert et al., 2020) and relapse (e.g., Lorimer et al., 2021) after CBT delivered in person, via internet or as a guided self-help intervention. Green et al. (2015), Saunders et al. (2016), Delgadillo et al. (2017), and Lorenzo-Luaces et al. (2017) applied ML techniques to predict which patients may be more or less likely to improve after stepped care interventions, where low and high intensity psychological treatments were accessed by patients sequentially. Other novel applications of prognostic ML models include the prediction of future dropout from treatment (Bennemann et al., 2022) and suicidal behaviors (Kessler et al., 2017).

Treatment selection models

Statistical models used to help clinicians to decide which of a number of treatment options to offer to a specific patient have been referred to as 'treatment selection models' (Cohen & DeRubeis, 2018), 'precision treatment rules' (Kessler & Luedtke, 2021), or 'stratified care models' in the context of selecting between interventions of differing levels of intensity and cost (Delgadillo et al., 2022). The general principle is to predict expected treatment outcomes for different treatment options and to select the one that maximizes improvement or minimizes adverse outcomes (i.e., side effects, dropout, suicidal risk, etc.). In the field of pharmacotherapy, for example, ML techniques have been used to develop models to select among different types of medications (e.g., Kim et al., 2019; Nunes et al., 2020). Similarly, ML techniques have been used to develop psychological treatment selection models to recommend CBT vs. interpersonal psychotherapy (Van Bronswijk et al., 2021), CBT vs. psychodynamic therapy (Schwartz et al., 2021), CBT vs. person-centered experiential therapy (Delgadillo & Gonzalez Salas Duhne, 2020), trauma-focused CBT vs. eye-movement desensitization and reprocessing (Deisenhofer et al., 2018), cognitive processing therapy vs. prolonged exposure (Keefe et al., 2018), and others. Novel developments in this area include the application of ML algorithms to decide which treatment techniques may be advantageous for specific patients (Webb et al., 2022) and how to optimally match patients to therapists in order to improve treatment response (Delgadillo, Rubel, et al., 2020).

Process models

A highly innovative development in the field of psychotherapy involves the analysis of multi-modal information (e.g., video, audio, text) from therapy sessions. The focus on smaller units within psychotherapy sessions such as the patients' and the therapists' speech-turns enables a more in-depth understanding of psychotherapy processes and outcomes. The words that clients and therapists use in psychotherapy sessions reflect their internal thoughts and emotions and reveal important information about their interaction. Several proof-of-concept studies have demonstrated the usefulness of NLP methods to analyze psychotherapy sessions (Aafjes-van Doorn et al., 2021). For example, Ewbank et al. (2020) trained a deep learning model to automatically classify therapist utterances (i.e., statements) from internet-enabled CBT transcripts for over 14,000 patients. These utterances were classed into 24 categories, some of which were associated with greater odds of treatment engagement and symptomatic improvement. In another study from the same research group (Ewbank et al., 2021), utterances from therapy transcripts of 34,000 patients were automatically classified by a deep learning algorithm into five categories, of which instances of 'change-talk' were associated with increased odds of symptomatic improvement. Tanana et al. (2021) applied NLP techniques to train algorithms capable of automatically detecting emotions expressed across a database of 97,497 utterances from psychotherapy transcripts, showing that qualitatively important features can be recognized by automated algorithms. In another study, Atzil-Slonim et al. (2021) analyzed transcripts from 873 psychotherapy sessions using topic models to investigate associations between topics that came up in sessions and their associations with alliance ruptures and patients' self-reported functioning and symptoms. Non-verbal speech-related characteristics such as intonation or vocalizations are directly associated with emotional changes (Bryan et al., 2018). The vocal channel provides a promising gateway for examining intra-personal and inter-personal emotional dynamics in psychotherapy and circumvents the need to rely on subjective measures (e.g., self-reports and clinician assessments) and can be assessed by objectively codified indices. Voice (more than other channels) lends itself to non-obtrusive measurement. Researchers today implement speech and voice-related measures to study psychotherapy processes (e.g., Tomicic et al., 2015). Vocal measures have been found helpful in identifying subtle yet clinically relevant changes in affective states in psychotherapy (e.g., Paz et al., 2021). For example, Flemotomos et al. (2022) used advanced voice recognition and NLP methods to develop an automated tool capable of processing raw recorded audio form motivational interviewing sessions, in order to rate the competence of therapists in delivering this intervention. These latter studies illustrate the potential for ML to generate insights about how therapy works and to develop automated tools to provide feedback to

therapists about their interactions, their adherence to effective techniques and how patients may be responding to these interactions.

Current state of the art and outlook

At the time of writing, the application of ML in the field of mental health is relatively new. Peer review, quality appraisal and reporting guidelines for ML studies (cited above) have only emerged in the last few years. Consequently, most of the studies in this space are at the early stages of the development pipeline for AI technologies, and only count with *level i* or *level ii* evidence (see Figure 2). For example, a recent review of 228 studies of clinical prediction models in psychiatry (Meehan et al., 2022) reported that only one-fifth of the clinical prediction models were validated in a statistically independent sample (*level iii* evidence). Even fewer studies have crossed the bridge from algorithmic model development towards clinical evaluations (*level iv* evidence) in feasibility studies and controlled trials (e.g., see Delgadillo et al., 2022; Lutz et al., 2022; Oliver et al., 2021). We foresee that in the coming decade the quality of evidence in this field will be enhanced through increased awareness of good practice guidelines to reduce known risks of bias and by greater adherence to guidelines for transparent reporting of ML studies.

To conclude, the selection of studies cited above shows how the use of ML has huge potential to transform mental health care, shifting this field to a new era where AI technologies can improve the precision of diagnosis, prognosis, treatment selection and delivery.

References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J. and Aafjes, M., 2021. A scoping review of machine learning in psychotherapy research. Psychotherapy Research, 31(1), 92-116.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D., 2006. The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. The Counseling Psychologist, 34(3), 341-382.

Archer, L., Snell, K.I., Ensor, J., Hudda, M.T., Collins, G.S. and Riley, R.D., 2021. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. Statistics in Medicine, 40(1), 133-146.

Atzil-Slonim, D., Juravski, D., Bar-Kalifa, E., Gilboa-Schechtman, E., Tuval-Mashiach, R., Shapira, N., & Goldberg, Y., 2021. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. Psychotherapy, 58(2), 324–339.

Baldwin, S. A., & Imel, Z. E., 2013. Therapist effects: Findings and methods. In M. J. Lambert (Ed.). Bergin and Garfield's handbook of psychotherapy and behavior change. In M. J. Lambert (Ed.), (6th ed., pp. 258–297). Hoboken, NJ: Wiley.

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W., 2022. Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. The British Journal of Psychiatry, 220(4), 192-201.

Blamire, P. A., 1996. The influence of relative sample size in training artificial neural networks. International Journal of Remote Sensing, 17(1), 223-230.

Bone, D., Lee, C. C., & Narayanan, S., 2014. Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. IEEE Transactions on Affective Computing, 5(2), 201–213.

Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199-231.

Bryan, C. J., Baucom, B. R., Crenshaw, A. O., Imel, Z., Atkins, D. C., Clemans, T. A., ... & Rudd, M. D., 2018. Associations of patient-rated emotional bond and vocally encoded emotional arousal among clinicians and acutely suicidal military personnel. Journal of Consulting and Clinical Psychology, 86(4), 372-383.

Bzdok, D., & Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: opportunities and challenges. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3(3), 223-230.

Catarino, A., Fawcett, J.M., Ewbank, M.P., Bateup, S., Cummins, R., Tablan, V. and Blackwell, A.D., 2022. Refining our understanding of depressive states and state transitions in response to cognitive behavioural therapy using latent Markov modelling. Psychological Medicine, 52(2), 332-341.

Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R. and Dwyer, D., 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry, 20(2), 154-170.

Cohen, Z.D., & DeRubeis, R.J., 2018. Treatment selection in depression. Annual Review of Clinical Psychology, 14, 209-236.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G., 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Journal of British Surgery, 102(3), 148-158.

Collins, G.S., Dhiman, P., Navarro, C.L.A., Ma, J., Hooft, L., Reitsma, J.B., Logullo, P., Beam, A.L., Peng, L., Van Calster, B. and van Smeden, M., 2021. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ open, 11(7), e048008.

Deisenhofer, A.K., Delgadillo, J., Rubel, J.A., Boehnke, J.R., Zimmermann, D., Schwartz, B., & Lutz, W., 2018. Individual treatment selection for patients with posttraumatic stress disorder. Depression and Anxiety, 35(6), 541-550.

Delgadillo, J., 2021. Machine learning: A primer for psychotherapy researchers. Psychotherapy Research, 31(1), 1-4.

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., Cohen, Z.D., DeRubeis, R.J. and Barkham, M., 2022. Stratified care vs stepped care for depression: A cluster randomized clinical trial. JAMA Psychiatry, 79(2), 101-108.

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D., 2017. Case complexity as a guide for psychological treatment selection. Journal of Consulting and Clinical Psychology, 85(9), 835–853.

Delgadillo, J., & Lutz, W., 2020. A development pathway towards precision mental health care. JAMA Psychiatry, 77(9), 889-890.

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.

Delgadillo, J., Rubel, J., & Barkham, M., 2020. Towards personalized allocation of patients to therapists. Journal of Consulting and Clinical Psychology, 88(9), 799–808.

DeRubeis, R.J., 2019. The history, current status, and possible future of precision mental health. Behaviour Research and Therapy, 123, 103506.

Dwyer, D., & Koutsouleris, N., 2022. Annual Research Review: Translational machine learning for child and adolescent psychiatry. Journal of Child Psychology and Psychiatry, 63(4) 421-443.

Dwyer, D.B., Falkai, P. and Koutsouleris, N., 2018. Machine learning approaches for clinical psychology and psychiatry. Annual Review of Clinical Psychology, 14, 91-118.

Efron, B., & Tibshirani, R., 1997. Improvements on cross-validation: The 632+ bootstrap method. Journal of the American Statistical Association, 92, 548–560.

Ewbank, M.P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A.J. and Blackwell, A.D., 2020. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. JAMA Psychiatry, 77(1), 35-43.

Ewbank, M.P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S. and Blackwell, A.D., 2021. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. Psychotherapy Research, 31(3), 300-312.

Fernandes, B.S., Williams, L.M., Steiner, J., Leboyer, M., Carvalho, A.F. and Berk, M., 2017. The new field of 'precision psychiatry'. BMC Medicine, 15, 80.

Flemotomos, N., Martinez, V.R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D.D., Gibson, J., Tanana, M.J., Georgiou, P., Van Epps, J., Lord, S.P., Hirsch, T., Imel, Z.E., Atkins, D.C., & Narayanan, S., 2022. Automated evaluation of psychotherapy skills using speech and language technologies. Behavior Research Methods, 54, 690–711.

Flygare, O., Enander, J., Andersson, E., Ljótsson, B., Ivanov, V.Z., Mataix-Cols, D. and Rück, C., 2020. Predictors of remission from body dysmorphic disorder after internet-delivered cognitive behavior therapy: a machine learning approach. BMC Psychiatry, 20, 247.

Green, S.A., Honeybourne, E., Chalkley, S.R., Poots, A.J., Woodcock, T., Price, G., Bell, D. and Green, J., 2015. A retrospective observational analysis to identify patient and treatment-related predictors of outcomes in a community mental health programme. BMJ Open, 5(5), e006103.

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., ... & 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.

Hastie, T., Friedman, J., & Tibshirani, R., 2009. The elements of statistical learning: data mining, inference, and prediction. New York: Springer.

Hilbert, K., Kunas, S.L., Lueken, U., Kathmann, N., Fydrich, T. and Fehm, L., 2020. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: A machine learning approach. Behaviour Research and Therapy, 124, 103530.

Indurkhya, N., & Damerau, F. J., 2010. Handbook of natural language processing. New York: Chapman and Hall/CRC.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C. and Wang, P., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. American Journal of Psychiatry, 167(7), 748-751.

Keefe, J.R., Wiltsey Stirman, S., Cohen, Z.D., DeRubeis, R.J., Smith, B.N. and Resick, P.A., 2018. In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. Depression and Anxiety, 35(4), 330-338.

Kessler, R.C., Hwang, I., Hoffmire, C.A., McCarthy, J.F., Petukhova, M.V., Rosellini, A.J., Sampson, N.A., Schneider, A.L., Bradley, P.A., Katz, I.R. and Thompson, C., 2017. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans Health Administration. International Journal of Methods in Psychiatric Research, 26(3), e1575.

Kessler, R.C., & Luedtke, A., 2021. Pragmatic Precision Psychiatry—A New Direction for Optimizing Treatment Selection. JAMA Psychiatry, 78(12), 1384-1390.

Kim, T.T., Dufour, S., Xu, C., Cohen, Z.D., Sylvia, L., Deckersbach, T., DeRubeis, R.J. and Nierenberg, A.A., 2019. Predictive modeling for response to lithium and quetiapine in bipolar disorder. Bipolar Disorders, 21(5), 428-436.

Kocak, B., Kus, E.A. and Kilickesmez, O., 2021. How to read and review papers on machine learning and artificial intelligence in radiology: a survival guide to key methodological concepts. European Radiology, 31(4), 1819-1830.

Laufer-Goldshtein, B., Talmon, R., & Gannot, S., 2018. Source counting and separation based on simplex analysis. IEEE Transactions on Signal Processing, 66(24), 6458–6473.

Li, X., Horaud, R., Girin, L., & Gannot, S., 2016. Voice activity detection based on statistical likelihood ratio with adaptive thresholding. In 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 1-5.

Lorimer, B., Delgadillo, J., Kellett, S. and Lawrence, J., 2021. Dynamic prediction and identification of cases at risk of relapse following completion of low-intensity cognitive behavioural therapy. Psychotherapy Research, 31(1), 19-32.

Lorenzo-Luaces, L., DeRubeis, R.J., van Straten, A. and Tiemens, B., 2017. A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining

multiple variables to inform risk-stratified stepped care models. Journal of Affective Disorders, 213, 78-85.

Lutz, W., Deisenhofer, A.-K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B., 2022. Prospective evaluation of a clinical decision support system in psychological therapy. Journal of Consulting and Clinical Psychology, 90(1), 90–106.

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W.B., Evans, C., Noble, R. and Iveson, S., 2005. Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. Journal of consulting and clinical psychology, 73(5), 904–913.

Lutz, W., Schwartz, B. and Delgadillo, J., 2022. Measurement-Based and Data-Informed Psychological Therapy. Annual Review of Clinical Psychology, 18. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Meehan, A.J., Lewis, S.J., Fazel, S., Fusar-Poli, P., Steyerberg, E.W., Stahl, D. and Danese, A., 2022. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Molecular Psychiatry. https://doi.org/10.1038/s41380-022-01528-4

Morgan, J., Daugherty, R., Hilchie, A. and Carey, B., 2003. Sample size and modeling accuracy of decision tree based data mining tools. Academy of Information and Management Sciences Journal, 6(2), 77-91.

Nunes, A., Ardau, R., Berghöfer, A., Bocchetta, A., Chillotti, C., Deiana, V., Garnham, J., Grof, E., Hajek, T., Manchia, M. and Müller-Oerlinghausen, B., 2020. Prediction of lithium response using clinical data. Acta Psychiatrica Scandinavica, 141(2), 131-141.

Oliver, D., Spada, G., Colling, C., Broadbent, M., Baldwin, H., Patel, R., Stewart, R., Stahl, D., Dobson, R., McGuire, P. and Fusar-Poli, P., 2021. Real-world implementation of precision psychiatry: transdiagnostic risk calculator for the automatic detection of individuals at-risk of psychosis. Schizophrenia Research, 227, 52-60.

Orru, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G. and Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience & Biobehavioral Reviews, 36(4), 1140-1152.

Paz, A., Rafaeli, E., Bar-Kalifa, E., Gilboa-Schectman, E., Gannot, S., Laufer-Goldshtein, B., ... & Atzil-Slonim, D., 2021. Intrapersonal and interpersonal vocal affect dynamics during psychotherapy. Journal of Consulting and Clinical Psychology, 89(3), 227-239.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., 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.

Riley, R.D., Snell, K.I., Ensor, J., Burke, D.L., Harrell Jr, F.E., Moons, K.G. and 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.

Rodriguez, J. D., Perez, A., & Lozano, J. A., 2009. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3), 569-575.

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.

Sajjadian, M., Lam, R.W., Milev, R., Rotzinger, S., Frey, B.N., Soares, C.N., Parikh, S.V., Foster, J.A., Turecki, G., Müller, D.J. and Strother, S.C., 2021. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychological Medicine, 51(16), 2742-2751.

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E.W. and Stahl, D., 2021. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. Schizophrenia bulletin, 47(2), 284-297.

Saunders, R., Cape, J., Fearon, P. and Pilling, S., 2016. Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. Journal of Affective Disorders, 197, 107-115.

Schwartz, B., Cohen, Z.D., Rubel, J.A., Zimmermann, D., Wittmann, W.W. and Lutz, W., 2021. Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. Psychotherapy Research, 31(1), 33-51.

Shatte, A.B., Hutchinson, D.M. and Teague, S.J., 2019. Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine, 49(9), 1426-1448.

Silva, R.F., Castro, E., Gupta, C.N., Cetin, M., Arbabshirani, M., Potluru, V.K., Plis, S.M. and Calhoun, V.D., 2014, September. The tenth annual MLSP competition: Schizophrenia classification challenge. In 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 1-6.

Simmonds-Buckley, M., Catarino, A., & Delgadillo, J., 2021. Depression subtypes and their response to cognitive behavioral therapy: A latent transition analysis. Depression and Anxiety, 38(9), 907-916.

Sounderajah, V., Ashrafian, H., Golub, R.M., Shetty, S., De Fauw, J., Hooft, L., Moons, K., Collins, G., Moher, D., Bossuyt, P.M. and Darzi, A., 2021. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open, 11(6), e047709.

Tanana, M.J., Soma, C.S., Kuo, P.B., Bertagnolli, N.M., Dembe, A., Pace, B.T., Srikumar, V., Atkins, D.C., & Imel, Z.E., 2021. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. Behavior Research Methods, 53(5), 2069-2082.

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.

Tomicic, A., Martinez, C., & Krause, M., 2015. The sound of change: A study of the psychotherapeutic process embodied in vocal expression. Laura Rice's ideas revisited. Psychotherapy research, 25(2), 263-276.

Torous, J., Onnela, J.P., & Keshavan, M., 2017. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. Translational psychiatry, 7(3), e1053-e1053.

Walter, M., Alizadeh, S., Jamalabadi, H., Lueken, U., Dannlowski, U., Walter, H., Olbrich, S., Colic, L., Kambeitz, J., Koutsouleris, N. and Hahn, T., 2019. Translational machine learning for psychiatric neuroimaging. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 91, 113-121.

Webb, C. A., Forgeard, M., Israel, E. S., Lovell-Smith, N., Beard, C., & Björgvinsson, T., 2022. Personalized prescriptions of therapeutic skills from patient characteristics: An ecological momentary assessment approach. Journal of Consulting and Clinical Psychology, 90(1), 51–60.

Whelan, R, Watts, R, Orr, CA, Althoff, RR, Artiges, E, Banaschewski, T, Barker, GJ, Bokde, AL, Buchel, C, Carvalho, FM, Conrod, PJ, Flor, H, Fauth-Buhler, M, Frouin, V, Gallinat, J, Gan, G, Gowland, P, Heinz, A, Ittermann, B, Lawrence, C, Mann, K, Martinot, JL, Nees, F, Ortiz, N, Paillere-Martinot, ML, Paus, T, Pausova, Z, Rietschel, M, Robbins, TW, Smolka, MN, Strohle, A, Schumann, G, Garavan, H, IMAGEN Consortium, 2014. Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature, 512, 185–193.

Wolff, R.F., Moons, K.G., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S. and PROBAST Group, 2019. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51-58.

Van Bronswijk, S.C., Bruijniks, S.J., Lorenzo-Luaces, L., Derubeis, R.J., Lemmens, L.H., Peeters, F.P., Huibers, M.J., 2021. Cross-trial prediction in psychotherapy: External validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. Psychotherapy Research, 31(1), 78-91.

Zhu, X., Goldberg, A.B., 2009. Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 3, 1–130.