The Journal of Child Psychology and Psychiatry

Journal of Child Psychology and Psychiatry 63:4 (2022), pp 421–443 doi:10.1111/jcpp.13545

Annual Research Review: Translational machine learning for child and adolescent psychiatry

Dominic Dwyer,[1,2,3] and Nikolaos Koutsouleris[1,4,5]

[1]Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Munich, Germany; [2]Orygen, The National Centre of Excellence in Youth Mental Health, Melbourne, Australia; [3]Centre for Youth Mental Health, University of Melbourne, Melbourne, Australia; [4]Max-Planck Institute of Psychiatry, Munich, Germany; [5]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Children and adolescents could benefit from the use of predictive tools that facilitate personalized diagnoses, prognoses, and treatment selection. Such tools have not yet been deployed using traditional statistical methods, potentially due to the limitations of the paradigm and the need to leverage large amounts of digital data. This review will suggest that a machine learning approach could address these challenges and is designed to introduce new readers to the background, methods, and results in the field. A rationale is first introduced followed by an outline of fundamental elements of machine learning approaches. To provide an overview of the use of the techniques in child and adolescent literature, a scoping review of broad trends is then presented. Selected studies are also highlighted in order to draw attention to research areas that are closest to translation and studies that exhibit a high degree of experimental innovation. Limitations to the research, and machine learning approaches generally, are outlined in the penultimate section highlighting issues related to sample sizes, validation, clinical utility, and ethical challenges. Finally, future directions are discussed that could enhance the possibility of clinical implementation and address specific questions relevant to the child and adolescent psychiatry. The review gives a broad overview of the machine learning paradigm in order to highlight the benefits of a shift in perspective towards practically oriented statistical solutions that aim to improve clinical care of children and adolescents. **Keywords:** Machine learning; ADHD; autism spectrum disorders; depression; psychosis; artificial intelligence.

Introduction

During childhood and adolescence, an estimated 15% of individuals will be diagnosed with autism spectrum disorder, attention deficit hyperactivity disorder, anxiety disorders, depression, and schizophrenia (Dalsgaard et al., 2020). Such diagnoses can be preceded by untreated periods of developmental disruption that are connected with poorer long-term outcomes (Dawson & Bernier, 2013; McGorry & Mei, 2018), and for some conditions, premorbid risk states characterized by subthreshold symptoms that impair functioning and can evolve into an illness during early adulthood (Fusar-Poli et al., 2013). These challenges have the capacity to impact on critical neurodevelopmental windows (Marin, 2016) in addition to altering social, psychological, and educational trajectories that increase risk of longer-term impairment. Facilitating the earliest detection of illness, providing accurate diagnoses, estimating prognostic courses, and predicting optimal treatments is thus critical to long-term health outcomes (Correll et al., 2018; Marin, 2016; McGorry, Ratheesh, & O'Donoghue, 2018).

Research in psychiatry and clinical psychology has provided a wealth of information based on groups of individuals that has, for example, facilitated earlier detection of disorders (McGorry & Mei, 2018), informed diagnostic categories (McPartland, Reichow, & Volkmar, 2012), generated treatment guidelines (Hollon et al., 2014), and highlighted potential biological mechanisms of illness (Insel & Cuthbert, 2015; Kapur, Phillips, & Insel, 2012). For example, an adolescent with depressive symptoms may be given a diagnosis of depression, they may be informed that individuals with the same diagnosis remit ~50% of the time without treatment (Vitiello, 2011), and they could be prescribed antidepressants or cognitive behavioral therapy (CBT) with a group-based remission possibility of 65% (Cox et al., 2014). This information is somewhat helpful, but in these cases where a diagnosis cannot accurately define an outcome or treatment plan (Hyman, 2010; Rødgaard, Jensen, Vergnes, Soulières, & Mottron, 2019; Vitiello, 2011), the patient and treating team would benefit from personalized approaches that deliver individualized risk estimates (Box 1).

Without accurate recommendations for each individual, the possibility of longer-term impairment increases along with the personal, societal, and economic cost of delayed diagnoses (Dawson & Bernier, 2013; Maenner et al., 2020), repeated clinical interactions, and trial-and-error treatment strategies (Chekroud et al., 2016). While there is a large body of research that has investigated associations between clinically relevant targets (e.g., poor outcomes) and psychological, social, and biological variables, the aims have largely been explanatory

Conflict of interest statement: No conflicts declared.

Box 1 Case example

Picture a clinician seeing a 16-year-old whose school performance has decreased, they are estranged from their friends, they are sad most days, and they are experiencing some mild paranoid symptoms regarding classmates talking about them behind their back. Major depressive disorder arises as a differential, but the clinician is concerned about the possibility of a psychosis prodrome. Group-based research suggests that mild paranoia is common in depressed adolescents with a history of trauma and within groups of individuals there is a low possibility of psychosis (Bird et al., 2021) and the clinician refers them to a depression clinic. However, the adolescent ultimately experiences a psychotic episode with subsequently poor social and symptomatic outcomes. On hearing of this outcome, the clinician wonders why years of research has not produced objective tools to turn the group-based risk estimate into something that could have been used for their individual patient.

(Yarkoni & Westfall, 2017) and none of the statistical models (e.g., regression equations) are routinely used in clinical care to deliver individualized risk that we are aware of. This is different to some other medical fields where simple equations can be used to predict outcomes on the basis of a small set of predictive variables, such as the Framingham score to estimate cardiovascular risk (e.g., using age, cholesterol, smoking status, and blood pressure) (Wilson et al., 1998).

Given the complexity and heterogeneity of psychiatric disorders it is perhaps understandable that simple risk calculators like the Framingham equation are not clinically used (Uddin, Wang, & Woodbury-Smith, 2019), especially due to the absence of objective biomarkers or a mechanistic understanding of psychiatric diagnoses (Abi-Dargham & Horga, 2016; Insel & Cuthbert, 2015; Kapur et al., 2012). However, what is increasingly difficult to understand is why an era characterized by digital data storage, high speed connectivity, large computational resources, and the widespread use of artificial intelligence in nonmedical fields (Jordan & Mitchell, 2015; LeCun, Bengio, & Hinton, 2015) has had a limited impact on the use of predictive algorithms in psychiatry to-date. This is especially the case with the emergence of massive repositories from primary care services in the form of electronic health records (EHR) and diagnostic data (e.g., magnetic resonance imaging; MRI) in addition to increasing collection of data from video, audio, smart phones, social media, clinical questionnaires, genomics, other -omics data, electrophysiology, neuroimaging, and many other sources (Russ et al., 2019b; Sim, 2019; Topol, 2019). These data sources could be used to harness the hypothesized biological, psychological, social, and environmental contributions to diagnoses, prognoses, and treatment predictions in psychiatry by reconsidering the dominant statistical paradigm and enhancing it with a translational machine learning approach.

P-value testing and group-based thinking

Psychiatric research mostly uses a frequentist, inferential statistical paradigm to design experiments and make conclusions about data (Bzdok, Altman, & Krzywinski, 2018; Bzdok, Engemann, & Thirion, 2020). Group-based effects are commonly the target (e.g., differences and associations) and conclusions are made on the basis of *p*-values indicating the probability of obtaining the result in the absence of a true effect, by calculating a statistical model under assumptions of null-hypothesis significance testing (Nuzzo, 2014). A traditional scientific method is assumed that aims to carefully design highly controlled experiments to sample from a population targeted for inference, select variables based on precise hypotheses, use statistical models to determine significance and interpret variables, and make conclusions on the basis of whether the results potentially occurred by chance or not (Naci & Ioannidis, 2015). By using such a paradigm, researchers make conclusions about whether a hypothesized effect occurs in the inferred population of cases—for example, whether trauma is associated with depression or whether CBT is appropriate to treat it. Within this paradigm, average risks are compared between groups.

Recently, the dominant analytic paradigm has been questioned in light of a reproducibility crisis that has highlighted multiple limitations (Ioannidis, 2005; Schooler, 2014). An established fact that has been revisited in this controversy is that *p*-values do not assess replicability or reproducibility (Goodman, 1992; Goodman, Fanelli, & Ioannidis, 2016; Nuzzo, 2014), which has contributed to a reevaluation of research practices, such as: the use of pretest probability estimates (Ioannidis, 2005), wider use of confidence intervals (Cumming, 2014), and preregistration of analysis plans (Nuzzo, 2014). However, even in the presence of such important changes, the remaining paradigm itself may still lend itself towards research that does not contribute to translational aims by producing results with clinically meaningless effect sizes (e.g., differences between groups; Abi-Dargham & Horga, 2016; Ioannidis, 2016), excessively controlling samples so that they no longer resemble real-world circumstances (Naci & Ioannidis, 2015), falsely identifying predictive variables on the basis of their significance (Lo, Chernoff, Zheng, & Lo, 2015), and making questionable inferences from group averages to individuals (Fisher, Medaglia, & Jeronimus, 2018). Ultimately, significance testing does not directly measure

generalizability from the sample under investigation to new cases, or predictive accuracy, and these are required if we want to practically use the statistical model to make a decision (Bzdok et al., 2020).

Machine learning for medicine

In order to use traditional group-based approaches for prediction at an individual level, statistical models are commonly validated in external samples and this has led to tools such as the Framingham risk equation mentioned above. However, due to limitations of these models, the medical field is investing in other techniques (Esteva et al., 2019; Rajkomar, Dean, & Kohane, 2019; Topol, 2019; Yu, Beam, & Kohane, 2018) that combine existing knowledge and practices with the complementary and overlapping paradigm of machine learning (Breiman, 2001; Bzdok et al., 2018).

The origins of this paradigm can be traced back to research using early computers to simulate the functioning of neurons, which was conducted by early interdisciplinary pioneers in psychology, neuroscience, and computer science, such as Frank Rosenblatt (Rosenblatt, 1958). Historically, the idea was to model basic neuronal operations by encoding a computer with statistical functions that could autonomously update their coefficients based on data input in order to classify new examples—in this case, the recognition of shape patterns (e.g., squares or triangles) from a sensor array that mimicked the retina of the eye. As such, the 'machine' was able to 'learn' from incoming 'features' of the shapes by using 'pattern recognition' in order to classify examples shown to it (see glossary in Table 1 for descriptions of machine learning terminology used in this review).

When employed by Rosenblatt and others, the use of statistics within machine learning was different from the dominant frequentist paradigm in psychology because it was focused on algorithmic approaches that facilitated learning from examples to achieve a practical goal of prediction rather than the use of statistical models and *p*-value testing. It was also different from standard practices of computing that involved programming preconceived rules into a computer to reach deterministic outcomes. In this way, machine learning was a semi-autonomous, probabilistic middle-road between statistics and computer science that created hopes of achieving human-level abilities (i.e., artificial intelligence). The central ideas can be distilled down to automatically selecting data, learning parameters, making limited assumptions, using simulation to assess and enhance performance, and making probabilistic predictions to drive specific decisions. Historically, the accuracy, generalizability to new cases, practical utility, and applicability to single examples (e.g., patients) was the main end goal rather than finding significant differences between group means or statistical associations (Bzdok et al., 2018; Bzdok & Ioannidis, 2019; Yarkoni & Westfall, 2017).

The machine learning paradigm fell in and out of favor in the ensuing years, but now is a part of our

Table 1 Glossary of machine learning terms used in this review

Accuracy	The fraction of correctly predicted cases in reference to all cases
Cross-validation	An internal validation resampling technique used to empirically assess the accuracy and generalizability of statistical models, usually for a specific outcome
Feature engineering	Modification of variables in order to enhance predictions (e.g., through data dimensionality reduction)
Feature selection	The selection of optimal variables without their modification
Generalizability	Algorithm performance on new data that can be assessed with internal validity (e.g., using cross-validation techniques) or external validity (e.g., validating the models on data from a different study, time period, or geographic location). Also includes the assessment of model bias towards certain dominant groups (e.g., Western European groups).
Hyperparameter	A modifiable setting of an algorithm that can be altered to obtain optimal prediction accuracy and generalizability
Information leakage	When information about test subjects is included in the training sample, usually by conducting procedures outside of a cross-validation cycle that necessitate the use of all subjects in the sample (e.g., selection of variables or control of covariates). Assessment of generalizability is undermined and accuracy estimates are invalid.
Overfitting	Fitting a model to noise and idiosyncratic attributes of a training sample resulting in low levels of test accuracy and lowering generalizability potential.
Sensitivity	The proportion of affected cases with a positive test result in reference to all affected cases.
Specificity	The proportion of nonaffected cases with a negative test result in reference to all nonaffected cases.
Supervised learning	When the target outcome is known for all cases and predictive algorithms seek to first classify the known outcomes and then to predict them in new cases.
Testing	Within an internal validation procedure such as cross-validation, testing is the application of trained models or pipelines without modification to held-out data that has not been used in the creation of the models.
Training	The statistical procedures usually conducted within a cross-validation routine that involve fitting a model to a dataset (e.g., for prediction).
Unsupervised learning	When the target outcomes or subgroups are not known and an exploratory approach is used to learn natural groupings of cases.
Labels	The predictive target used in supervised learning, that is, assigned to each case, such as diagnoses or prognostic outcomes.

daily lives because increases in computing power and algorithm advances led to performance gains that exceeded the use of traditional statistical approaches or rule-based, deterministic programming across a diverse array of fields (Jordan & Mitchell, 2015; Topol, 2019). Within this context, an old question has been revived about whether the same techniques can be used in medical contexts (Wilson et al., 1998) to improve predictions and create computer algorithms for diagnoses, prognoses, and treatment selection purposes (for an early example of computer-aided decision making in pediatrics see Barness, Tunnessen, Worley, Simmons, & Ringe, 1974). Across medicine, the use of machine learning techniques now exhibits exponential growth as the methods are tested for their ability to assist in decision-making across the lifespan in diverse specialties (Topol, 2019) with promising results (Esteva et al., 2019; Rajkomar et al., 2019; Topol, 2019; Yu et al., 2018).

The machine learning paradigm in medicine is particularly well suited for digital data, that is, difficult to analyze with simple regression equations or decision rules—for example, EHRs, medical images (e.g., CAT, MRI, cellular pathology, dermatology), or genomics (Esteva et al., 2019; Hosny, Parmar, Quackenbush, Schwartz, & Aerts, 2018; Rajkomar et al., 2019; Topol, 2019; Yu et al., 2018). However, even in cases where simple clinical data is collected (e.g., from a questionnaire) machine learning could assist in finding maximally predictive patterns, especially when existing knowledge is not sufficient to derive a clinically useful regression equation, the predictive target is new, or mechanistic understanding is insufficient. As such, it is a paradigm that is well-suited to the field of clinical psychology and psychiatry because it is not only a field with limited mechanistic insight (Abi-Dargham & Horga, 2016; Insel & Cuthbert, 2015; Kapur et al., 2012), but it is also one that attempts to find patterns in digital data sources (e.g., imaging or genetics) in order to ultimately assist with clinical care.

Translational machine learning fundamentals

The following section introduces key concepts and methods in machine learning in order to define machine learning objectives, introduce common techniques, discuss the use of statistical pipelines, emphasize the importance of algorithm optimization, and highlight the critical importance of validation and generalizability. For further information, there are a number of primers and reviews for medicine generally (Esteva et al., 2019; Rajkomar et al., 2019; Topol, 2019; Yu et al., 2018) and specifically for psychiatry (Bzdok & Ioannidis, 2019; Bzdok & Meyer-Lindenberg, 2018; Dwyer, Falkai, & Koutsouleris, 2018; Russ et al., 2019a; Rutledge, Chekroud, & Huys, 2019; Shatte, Hutchinson, & Teague, 2019), psychology (Yarkoni & Westfall, 2017), neurodevelopmental disorders (Uddin et al., 2019), and radiology (Hosny et al., 2018; Moore, Slonimsky, Long, Sze, & Iyer, 2019). There are also textbooks that expand on the topics covered below (Hastie, Tibshirani, & Friedman, 2009; James, Witten, Hastie, & Tibshirani, 2015).

Machine learning objectives

Data can be analyzed based on two main machine learning objectives. Supervised learning is the focus of this review and is when the 'labels' (e.g., the assignment of a predictive target to a case, such as a specific diagnosis or prognosis) are known and algorithms are optimized to find patterns in the data that separate cases; it is called 'supervised' learning because the labels are provided like a teacher would supervise students. Conversely, unsupervised learning, which is discussed as a future direction at the end of this article, is when labels are unknown and the algorithms are used to autonomously find patterns that separate the cases into clusters—for example, the simplest approach is the k-means method. There are also other techniques that will not be a focus of this review because they are currently infrequently used in psychiatry, such as semisupervised learning and reinforcement learning (Jordan & Mitchell, 2015).

Feature engineering

A traditional approach to building a predictive statistical tool (e.g., using logistic regression or Cox models) is to define a restricted set of variables based on hypotheses where the number of predictors is much less than the number of observations (e.g., 20 observations per variable; Ogundimu, Altman, & Collins, 2016). The variables are entered into a table and a linear model is fit to the data. Sometimes interaction or polynomial terms are added in order to better model the fit between the data and the outcome (e.g., transition to psychosis) or data are transformed in order to satisfy statistical assumptions (e.g., log transformation). This approach has produced such tools as the Framingham risk score for cardiovascular disease as described above.

In addition to hypothesis-driven techniques that define variable subsets, exploratory techniques can be used. For example, a simple technique, that is used across statistics and machine learning fields is to select subsets of the data using procedures that are similar to step-wise regression (Chandrashekar & Sahin, 2014; Kohavi & John, 1997). Another approach, which is essential to many machine learning studies, is to reduce the dimensionality of the data using techniques that preserve the variance (e.g., the differences between cases), such as principal components analysis (PCA; Hotelling, 1933), exploratory matrix factorization techniques (Lee &

Seung, 1999), and other more specialized dimensionality reduction methods (Zhang, Yan, & Lades, 1997). Such techniques are shared across different fields, including psychology, engineering, and computer science. Within a machine learning context, when the explanatory variables are modified it is called 'feature engineering', including standardization (e.g., Z-scoring), adding interaction terms, performing subset selection, dimensionality reduction, or any other technique that modifies the original data or creates new variables.

Classification and regression

While reducing the number of predictors using selection or reduction is effective, a defining feature of machine learning is that algorithms have been created that directly address the core limitations of traditional methods (Hastie et al., 2009). For example, a problem with standard regression techniques (e.g., least squares regression as a simple example) is that as the number of predictors approaches the number of observations the model will perfectly fit to the sample (Hastie et al., 2009). Another way to describe this is by stating that there will be no 'bias', defined by a difference between the estimated values from the model fit and the true values, but high model 'variance' because minor differences in the sample will result in changes to the coefficients of the model (Figure 1). This 'overfitting' to the sample results in poor predictions in new cases and rules within traditional statistics attempt to prevent it by restricting the p to n ratio as described above (Ogundimu et al., 2016).

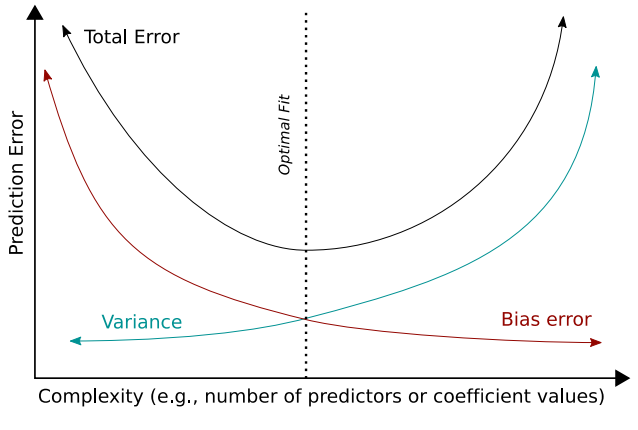

Figure 1 Machine learning approaches aim to balance bias and variance associated with the prediction error. When bias is high (i.e., the model does not fit the data) then the variance (i.e., the stability of the model to changes of input data) will be low. When bias is low (i.e., when the model perfectly fits the data), then the variance will be high because minor changes in the input data will result in changes to the model or algorithm. Algorithms attempt to find the optimal fit to balance these two extremes without constraining the entered data using a variety of techniques, such as automatic feature selection, dimensionality reduction, or regularization to constrain the coefficients

Another approach is to directly address the bias-variance problem within an algorithm. By introducing a mathematical constraint on how the coefficients of the regression are calculated, analyses can be conducted where p variables approximate n cases or even when $p > n$. This constraint is called regularization, which can decrease the coefficients of variables in an equation in order to automatically attenuate the possibility of overfitting (Hastie et al., 2009; Yarkoni & Westfall, 2017). The most common forms of regularization for regression are the ridge and the lasso approaches (Tibshirani, 1996), with the latter being able to shrink coefficients to exactly zero in order to act as an automatic 'feature selection' technique that is also highly interpretable. The amount of this shrinkage is defined by a 'hyperparameter' (i.e., a parameter that modifies other parameters, which in this case are the coefficients) that can be set by the experimenter or automatically detected in order to produce a model that balances bias error and variance.

Regularization, and the use of hyperparameters to balance the bias-variance tradeoff, are also used by many other algorithms. For supervised learning, these approaches are commonly classification methods, such as regularized logistic regression (L2- or L1-regularized versions) or support vector machine (SVM) techniques that were developed simultaneously in the fields of computer science and engineering (Boser, Guyon, & Vapnik, 1992). The SVM is especially important because it has been highly effective in previous research and is widely used in psychiatry. It aims to maximize a margin between groups in order to define a boundary (e.g., between good and poor outcomes) on the basis of individual cases called 'support vectors' (Figure 2). The strength of the SVM is that the margin can be modified by a hyperparameter that has the effect of allowing more or less misclassification of weighted cases in addition to modifying the coefficients (usually with L2-regularization). In similarity to the ridge or lasso regression, this has the effect of balancing the bias-variance tradeoff and has been very effective.

There are a wide range of machine learning algorithms that have been developed from multiple intersecting fields of statistics, engineering, and computer science. Some of these are similar to the regularized regression above in that they are developments to existing statistical techniques (e.g., decision trees developing into random forest algorithms), others have grown over time within specialized machine learning fields (e.g., neural network algorithms), and there are methods that have developed based on intersections between fields (e.g., boosting techniques where multiple decisions are combined together). Two subfields that will be important for the future of psychiatry are deep learning (Box 2) and computer perception including

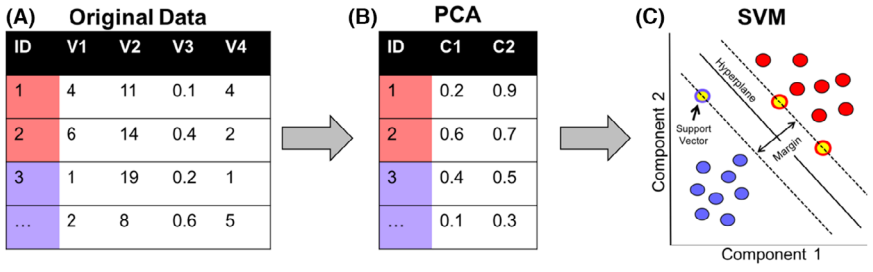

Figure 2 A simple analysis chain in machine learning. (A) Data is acquired with multiple variables (V1–V4); (B) the data are reduced using PCA into components with their associated weights for each individual depicted here (i.e., how much an individual fits the pattern represented by the component); (C) the components are forwarded to a SVM algorithm that finds a decision boundary separating the classes using an iterative learning approach. When the pipeline is embedded in a cross-validation framework (Figure 3), the number of components retained in the PCA and additional settings of the SVM can be optimized to balance bias and variance

computer vision and natural language processing (Box 3).

Regardless of the specific method, machine learning techniques are unified by their use of computers to find maximally predictive patterns, usually within large amounts of data, and directly addressing the bias-variance tradeoff rather than accepting the limitations of traditional methods. They are useful because rather than constraining the experimental design and data many of the techniques attempt to constrain algorithms and use iterative optimization algorithms to find the most predictive solutions. As will be outlined later in this review, there are limitations to this approach, but thus far, the additional tools have been effective to make sense of large bodies of data for the purposes of prediction.

Pipelines

Having a toolkit of feature engineering approaches and pattern recognition algorithms is helpful for predictions, but the challenge is often in combining approaches with data preparation methods to achieve a predictive goal. Machine learning is mostly still guided by expert knowledge regarding the data domain that is being investigated (except for deep learning, see Box 2), which means that specific preparations of the data are required and these are usually linked together in analysis chains (Figure 2). For example, after devising a hypothesis and selecting initial data, the first link in a machine learning chain is often to clean, scale, impute, transform, or perform other basic operations (as would be done in standard analyses before analysis). The data may then be forwarded to a dimensionality reduction algorithm (e.g., PCA) before being analyzed with a more specific classification algorithm to determine the classes of each individual (e.g., an SVM).

A further addition to machine learning pipelines is the ability to combine the predictions of individual models together in ensemble learning procedures before calculating a final decision (Polikar, 2006). Ensemble learning differs from traditional approaches where one statistical model is used to predict an outcome (e.g., the Framingham risk equation) and can improve predictive accuracy by increasing diversity of predictions—for example, models from different statistical algorithms or data types can be combined together into committees that ultimately decide on a prediction, which is conceptually similar to the way that committees of individuals like medical experts come to a decision. These techniques are used to enhance the stability and accuracy of predictions.

Within the context of machine learning pipelines, it is important note that a fundamental attribute of machine learning is the ability to optimize any step in a pipeline (or the entire pipeline) for optimal predictive accuracy. For example, instead of reducing the data to a number of dimensions using PCA that is defined by statistical rules-of-thumb (e.g., knee-point detection), the number of components can be automatically chosen based on their ability to predict outcomes in individuals. As described above, for machine learning-specific algorithms (e.g., SVM), hyperparameters can also be tuned to ultimately modify the number of features selected, degree of error allowed, or the amount of nonlinearity (Boser et al., 1992; Cortes & Vapnik, 1995). While some of these techniques are also used in traditional statistics, the learning field is more focused on this this capacity to autonomously and flexibly optimize elements of a pipeline to maximize predictive capacity (Bzdok & Ioannidis, 2019; Yarkoni & Westfall, 2017).

Cross-validation

Given the power of machine learning approaches to find optimal solutions in data using advanced algorithms that consist of multiple pipeline steps, the possibility of severe overfitting is high. As such, empirically testing and reporting the performance of algorithm pipelines in unseen tests cases and contexts is a fundamental component (Poldrack, Huckins, & Varoquaux, 2020; Varoquaux, 2018; Varoquaux et al., 2017).

Machine learning approaches assess, and ultimately aim to enhance, generalizability by using

Box 2 Deep learning

One subfield of machine learning that is often used to achieve the extreme of a hypothesis-free, unbiased feature engineering approach is deep learning. This family of techniques is particularly suited to problems where there is no existing knowledge or when re-evaluation of feature spaces is required (Esteva et al., 2019; LeCun et al., 2015). At a basic level, analysis chains of interconnected equations decompose raw data into layers of abstract features that together identify patterns using parts-based representations (e.g., parts of a face or tumor). For example, the first layer of an image analysis chain may identify simple lines, the second layer may be shapes, and the third layer could be more complete objects. The number of layers indicates the depth of the learning process (i.e., at what level of abstraction that the machine learns to achieve the goal). The statistical innovation in deep learning was to interconnect equations across layers of abstract data representations and allow the manipulation of weights associated with each connection (Esteva et al., 2019), which implies that the parameters of each linear equation are dependent on many others and the entire network can be trained to learn patterns. Thousands of hyperparameters can be tuned in this way given enough computational power, which offers a very high degree of autonomy that can be harnessed effectively to increase predictive accuracy but also increases the chances of finding spurious results. Deep learning approaches can package easily accessible pipelines elements that are usually conducted within conventional machine learning analyses (see Figure 2) into cohesive predictive framework (e.g., feature selection, dimensionality reduction, or ensemble building). In child and adolescent psychiatry, deep learning is currently being trialed in such fields as: neuroimaging based on the hypothesis that existing data preprocessing techniques (e.g., structural or functional data preparation) are restricting predictive accuracy (Riaz, Asad, Alonso, & Slabaugh, 2020), for EHR where there are huge corpuses of data with limited structure (Miotto, Li, Kidd, & Dudley, 2016), and for speech (Eni et al., 2020) and video (Li et al., 2020) recordings.

Box 3 Machine vision and natural language processing

A long-standing aim in artificial intelligence fields has been to create tools that can be used to interpret visual scenes (e.g., photos and video) and to process language (e.g., spoken and written) (Rosenblatt, 1958). Such tools are important for the future of psychiatry because diagnoses often involve speech and behavioral assessments (e.g., in autism spectrum diagnoses). Historically, the methods used in these fields are no different to the simple analysis chains described in the main text of this article (Turk & Pentland, 1991), but more recently deep learning has become dominant due to the availability of extremely large databases (e.g., YouTube or images on the web) and huge computing resources (LeCun et al., 2015). Early examples in the child and adolescent psychiatry field are where trained models (e.g., from YouTube) have been used to identify behavioral differences in autism using raw video files (Cook et al., 2019; Li et al., 2020; Preetham et al., 2017) and advances in speech recognition have been trialed to predict psychosis onset (Corcoran et al., 2018). Such advances are likely to continue as the field broadens and tools become easier to use, but are currently limited.

cross-validation approaches (i.e., 'internal validation') that leverage computer resources to simulate the circumstance of constructing a predictive algorithm and applying it to new data (Figure 3). A basic form of cross-validation could involve splitting a sample into two segments and applying a model built in one part to the other to assess its accuracy. However, in practice the methods employ repeated resampling procedures to test multiple subsets of data in order to increase the accuracy of the validation (Varoquaux, 2018; Varoquaux et al., 2017). It is important to note that cross-validation differs from other resampling techniques, such as bootstrapping, because it involves validating the statistical model or pipeline in a held-out sample (e.g., predicting outcome in individuals that are not included in the creation of the model) rather than repetitively applying the algorithm to the same sample with minor variations (e.g., resampling with replacement).

The simplest, but least accurate (Varoquaux, 2018), cross-validation technique is to leave one test subject out, fit an algorithm with the remaining data (called 'training'), and then apply the algorithm without modification to the left-out test case in a process called 'leave-one-out' cross-validation. There are multiple variations of such strategies that mostly consist of variations to the number of cases in each left-out subset of cases (called a 'fold'). The commonest approach is the k-fold where the data is first randomly separated into a predefined number of folds (e.g., 5 or 10). Each fold is then used as a testing sample while the rest of the data is used for training an algorithm and the average accuracy or other performance measure is calculated across test folds (Figure 3). Other variations involve leaving out a specific group as a fold instead of a random subset of the data, such as a hospital site in a multisite consortium, in order to determine if the models generalize to members of such groups. Cross-

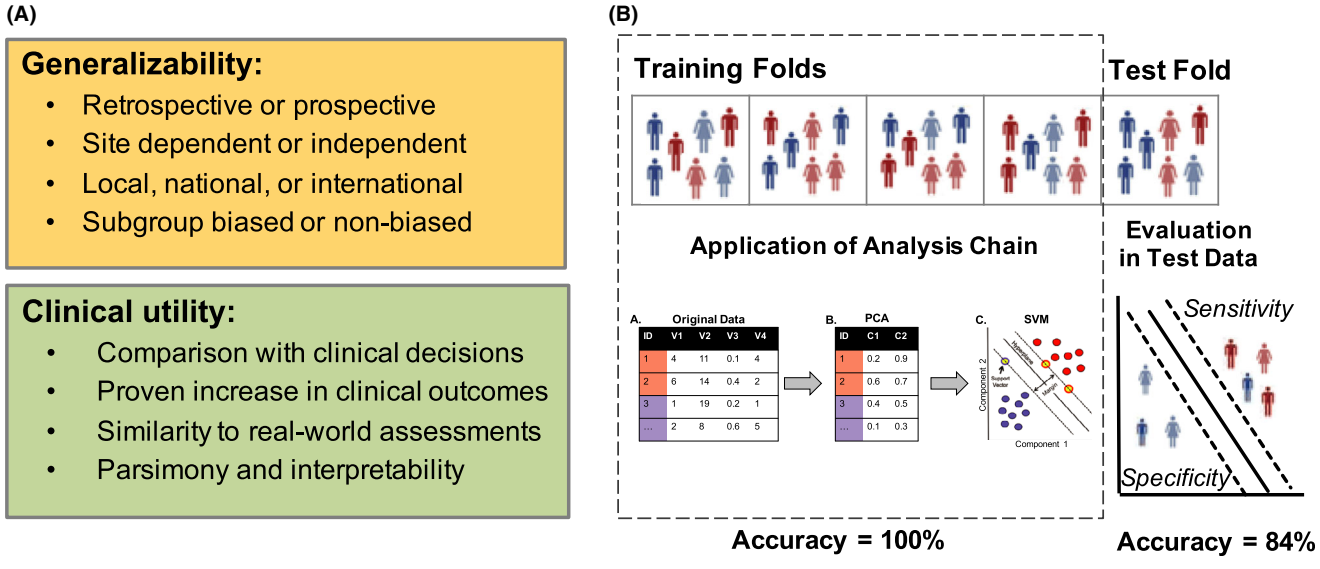

Figure 3 Generalizability and clinical utility of translational machine learning. (A) Translational machine learning pipelines should be evaluated based on their generalizability and clinical utility. (B) Cross-validation can be used to simulate generalizability in order to build and optimize models that are most likely to generalize. A k-fold cross-validation is depicted that involves dividing the sample into test folds. One test fold is held out, an analysis chain (e.g., Figure 2) is applied to the remainder of the sample and then the analysis chain is applied without modification to the held-out individuals to test whether it successfully predicts the outcome in each individual

validation techniques are highly flexible in this way and can assess multiple forms of generalizability (e.g., generalization to different regions, cultures, or genders) with the only rule being that individuals within the test folds must not be involved in the creation of the models—because this would undermine the simulation of testing accuracy in new data.

Nested cross-validation especially has been highly effective in optimizing algorithm pipelines and hyperparameters for maximal generalizability, while reducing the possibility of overfitting (i.e., balancing the bias-variance tradeoff Filzmoser, Liebmann, & Varmuza, 2009; Koutsouleris et al., 2018). This technique involves embedding a cross-validation routine inside another cross-validation cycle. Doing so allows models to be optimized on the test data of an inner cycle in order to learn the most generalizable patterns in unseen cases before applying these test-optimized analysis chains to the completely held-out individuals of the outer cross-validation loop. In general, nested cross-validation should be a standard for the field unless massive samples are used that can adequately sample the real population (e.g., for deep learning using EHR).

External validation and clinical utility

Internal cross-validation is necessary, but it does not obviate the need to further test generalizability if an algorithm will be used clinically. Generalizability can be assessed in terms of whether the models are accurate within single sites and samples (e.g., hospitals) or across multiple sites at a local (e.g., same city), national, or international levels in diverse samples that are representative of the demographic and clinical heterogeneity of the populations targeted by the statistical tool (Figure 3). As outlined in other reviews (Dwyer et al., 2018), a hierarchy of generalizability needs to be considered for any study that claims to have translational potential containing: internal validation within one site, internal validation with multiple sites, leave-site analyses within one study, external validation in a separate study, and ultimately prospective validation. It is also increasingly becoming important to actively test for model biases by validating against samples from different countries, ethnic groups, and genders. The current standard for strong generalizability claims is to validate the algorithms in a separate sample that has been collected as part of a different study (i.e., to provide evidence of external validation), but regulatory approvals for the use of medical algorithms are likely to be based on prospective validation procedures involving multisite randomized clinical trials that assess generalizability and clinical utility.

Clinical utility is a relatively neglected element of translational machine learning that considers whether the tool could practically be implemented in clinical care (e.g., whether the technology is available), whether it adds value to existing practices, how much value it adds given the incidence of the condition, whether the results can be interpreted by the clinical team, and the cost-benefit ratio of implementing the tool in care (Fusar-Poli, Hijazi, Stahl, & Steyerberg, 2018; Poldrack et al., 2020). In addition to the accuracy of an algorithm, its clinical utility can first be determined using net benefit analyses based on the relative harms of false-positive or false-negative results when considering the incidence of the disorder (Fusar-Poli et al., 2018).

However, even if a tool demonstrates a high accuracy, is generalizable, has a high degree of benefit to routine procedures, and is cost effective, recent attempts at deployment in medicine emphasize the need to work with clinical teams in order to determine how to integrate it into clinical workflows that are often variable between treating teams and across time (Beede et al., 2020; Gulshan et al., 2016). One element of the need for high clinical utility is the need for trustworthy predictions from interpretable models that evidence how the predictions were made. As such, a field of considerable interest currently is interpretable machine learning because it uses additional statistical techniques in order to make the sometimes opaque model predictions more transparent at an individual patient level (e.g., demonstrating why the patient was predicted to have a good or poor outcome on the basis of the data; Molnar, 2020; Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). Such concerns are expected to become more relevant as algorithms are clinically deployed, leading to calls for human-centered learning systems (Beede et al., 2020; Gulshan et al., 2016).

Analysis example

An example of an analysis pipeline that has been useful in psychiatry for neuroimaging may involve such steps as entering brain maps into a chain involving scaling, PCA, and then the use of an SVM to predict an outcome (Koutsouleris et al., 2021). This pipeline would be embedded within a nested cross-validation design consisting of an inner training-testing cycle that optimizes hyperparameters for prediction in held-out test samples. For each fold in the cross-validation cycles, the entire analysis pipeline is conducted on the training data, the best models are chosen on the basis of their ability to generalize to new cases, and then these models are applied without modification to the held-out individuals in the test folds in order to ultimately assess model performance accuracy. Once the training process is completed, the models can then be flexibly applied to other data, built into an online prognostic tool (www.proniapredictors.eu), or transferred to other sites for the further assessment of generalizability or the use in other conditions.

Machine learning in child & adolescent psychiatry

To provide a broad overview of machine learning research and to identify focus articles for further discussion, four scoping reviews in child and adolescent psychiatry were conducted (PRISMA guidelines; PubMed/Web of Science; Appendix S1) focusing on autism, attention-deficit hyperactivity disorder, early psychosis and psychosis risk, and depression (Table 2). Abstracts were excluded if they did not refer to machine learning, if they did not include a specific machine learning statistical technique, if they did not specify the data type that was used in the study, if the target was not specified, if they were not specifically focused on children or adolescents, or if a mental illness diagnosis or outcome was not specified. A total of 3,095 studies were screened of which 441 were retained after filtering (250, autism; 108, ADHD; 37, psychosis; 46, depression). Long-form conference abstracts from engineering fields (e.g., IEEE conferences) were retained in order to include pilot studies using experimental cutting-edge methods, with a total of 78 long-form conference abstracts included (51 of all autism studies (20%); 19 of all ADHD studies (18%); 7 of all depression studies (15%); 1 psychosis study (2%)). The clinical targets of the machine learning analyses (e.g., diagnoses, prognoses, or treatment selection) and the data type investigated (e.g., questionnaires, video, or MRI) were quantified for each study (Table 2). Selected studies in the domains of diagnosis, prognosis, and treatment selection were then discussed in order to provide example of the use of machine learning and also to assess the potential for the results to be clinically translated.

Broad research trends

Results demonstrated expected exponential increases in publication reflective of the increasing popularity of the field (Figure 4; Table 2). The highest number of publications was identified for autism. Diagnoses were a particular focus in autism and ADHD (74%), whereas early detection, prognoses, and symptom characterization were the main focus of psychosis (46%) and depression (44%). A wide range of data types were used, but with a particular focus on questionnaires, neuroimaging, EEG, and video and motion tracking in autism specifically.

Diagnosis

Child and adolescent diagnostic assessments can be laborious and highly specialized, which has the potential to lead to treatment delays during critical developmental windows and the possibility of misdiagnosis (Abbas, Garberson, Liu-Mayo, Glover, & Wall, 2020). Diagnostic machine learning techniques have thus been proposed for almost 30 years (Cohen, Sudhalter, Landon-Jimenez, & Keogh, 1993) with the aim to reduce the assessment burden (e.g., reduce time to make a diagnosis), especially in the fields of autism and ADHD (Table 2). While neuroimaging has generated the most research, questionnaire-based assessments have offered more possibility of translation because of higher evidence of generalizability. An exciting array of research also suggests that the future of autism spectrum diagnosis could involve the use of diverse digital data.

Table 2 Types of machine learning studies divided by diagnosis

	Autism	ADHD	Psychosis	Depression	Chi Sq.(df)	*p**	Phi
Clinical target[a]							
Early detection	18 (7.2)	0 (0.0)	8 (21.6)	5 (10.9)	21.27 (8)	<.001	0.22
Diagnoses	185 (74.0)	91 (84.3)	13 (35.1)	12 (26.1)	73.05 (8)	<.001	0.41
Differential diagnoses	10 (4.0)	9 (8.3)	0 (0.0)	2 (4.3)	5.22 (8)	n.s	0.11
Severity/Symptoms	25 (10.0)	3 (2.8)	1 (2.7)	20 (43.5)	59.35 (8)	<.001	0.37
Prognoses	1 (0.4)	0 (0.0)	17 (45.9)	4 (8.7)	149.06 (8)	<.001	0.58
Treatment	7 (2.8)	3 (2.8)	3 (8.1)	2 (4.3)	3.02 (8)	n.s	0.08
Subgroup definition	13 (5.2)	2 (1.9)	0 (0.0)	3 (6.5)	4.44 (8)	n.s	0.1
Data Modality[a]							
EHR	7 (2.8)	1 (0.9)	3 (8.1)	3 (6.5)	6.50 (8)	n.s	0.12
Questionnaires	33 (13.2)	12 (11.1)	9 (24.3)	18 (39.1)	23.18 (8)	<.001	0.23
Cognitive Testing	3 (1.2)	9 (8.3)	6 (16.2)	1 (2.2)	23.34 (8)	<.001	0.23
Neuroimaging	89 (35.6)	60 (55.6)	24 (64.9)	15 (32.6)	21.80 (8)	<.001	0.22
EEG	22 (8.8)	20 (18.5)	1 (2.7)	1 (2.2)	14.46 (8)	<.001	0.18
Omics & Biochemistry	27 (10.8)	4 (3.7)	3 (8.1)	2 (4.3)	6.08 (8)	n.s	0.12
Speech	15 (6.0)	0 (0.0)	1 (2.7)	3 (6.5)	7.38 (8)	n.s	0.13
Video & Tracking	46 (18.4)	2 (1.9)	0 (0.0)	2 (4.3)	29.04 (8)	<.001	0.26
Wearables	9 (3.6)	6 (5.6)	0 (0.0)	2 (4.3)	2.40 (8)	n.s	0.07
Computer games	4 (1.6)	3 (2.8)	0 (0.0)	0 (0.0)	2.32 (8)	n.s	0.07
Virtual Reality	4 (1.6)	1 (0.9)	0 (0.0)	0 (0.0)	1.48 (8)	n.s	0.06
Robot Interactions	9 (3.6)	0 (0.0)	0 (0.0)	0 (0.0)	7.02 (8)	n.s	0.13
Social Networks	1 (0.4)	0 (0.0)	1 (2.7)	2 (4.3)	9.09 (8)	n.s	0.14
External Validation							
External validation	13 (5.2)	5 (4.6)	3 (8.1)	0 (0.0)	3.32 (8)	n.s	0.09

Phi, Phi coefficient of effect size for nonparameteric tests; EHR, electronic heath records; EEG, electroencephalogram; Neuroimaging includes all MRI modalities in addition to functional near infrared spectroscopy (fNIRS)
*Only *p*-values significant at a false-discovery rate of $p < .05$ shown.
[a]All Clinical Targets and Data Modalities were counted resulting in studies being counted multiple times for each category.

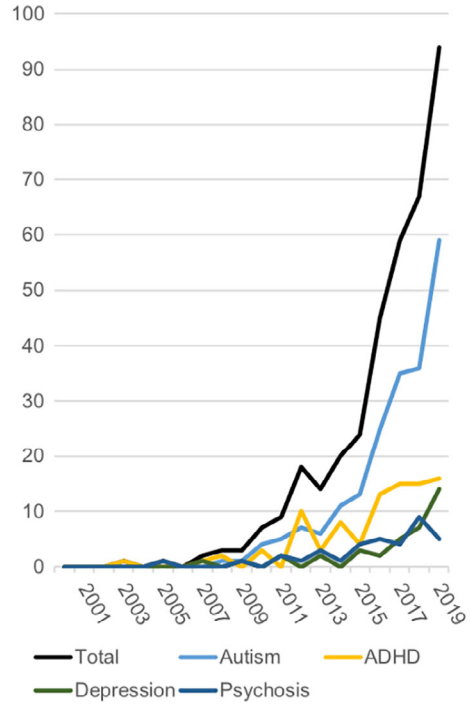

Figure 4 Machine learning publications in child and adolescent psychiatry. A main exponential trend is depicted for all diagnostic groups, that is, driven by studies in autism. The most active period has been in the last 5 years

Early machine learning approaches aimed at providing clinical support by learning diagnostic patterns in autism spectrum disorder from diagnostic assessment batteries that assess atypical behavioral and socio-emotional patterns (Cohen et al., 1993), which has been more recently revived (Wall, Kosmicki, Deluca, Harstad, & Fusaro, 2012) and applied to screening questionnaires (Achenie et al., 2019; Bone et al., 2016; Duda, Ma, Haber, & Wall, 2016a; Maenner, Yeargin-Allsopp, Van Naarden Braun, Christensen, & Schieve, 2016). Generalizability has been assessed in terms of multisite cross-validation (Abbas, Garberson, Glover, & Wall, 2018), external validation in other samples (Duda, Kosmicki, & Wall, 2014; Tariq et al., 2018), and differential diagnostic validation (Duda et al., 2016a), in addition to the study of gender, ethnicity, and education biases (Achenie et al., 2019) and the effects of cultural context (Tariq et al., 2019; Wingfield et al., 2020). Early detection has also been a focus (Abbas, Garberson, Glover, & Wall, 2017), which is also facilitated by the construction of parsimonious parental questionnaires using machine learning techniques (Abbas et al., 2018; Ben-Sasson, Robins, & Yom-Tov, 2018; Duda et al., 2016a).

On the basis of external validation and generalizability demonstrated in previous studies (Abbas et al., 2017, 2018, 2020; Duda, Haber, Daniels, & Wall, 2017; Duda et al., 2014, 2016a; Kosmicki, Sochat, Duda, & Wall, 2015; Tariq et al., 2018, 2019; Wall et al., 2012; Washington et al., 2020), a diagnostic assessment tool for autism have been developed that has recently received marketing approval from the Federal Drug Administration

(FDA) through the De Novo premarket review pathway for low- to moderate-risk devices (called Cognoa; www.cognoa.com). This approval was obtained on the basis of a (yet unpublished) multisite, prospective, double-blinded, cohort study conducted at 14 sites within the United States. The study involved the use of machine learning in combination with a mobile app that collects data from questionnaire items asked to the caregivers, manages uploaded video rated by manufacturer-trained and certified specialists, and provides a health care provider portal in order for clinicians to answer further questions in addition to providing reports. In a sample of 425 patients (18-months to 5 years) with concerns for developmental delay, the software based medical device was able to match diagnoses made by specialists for 32% of the sample and demonstrated an accuracy of 89% (98% sensitivity, 79% specificity). This approval demonstrates a first translation of machine learning models with questionnaire measures to assist in early diagnosis, which has the potential to improve outcomes and to access diagnostic procedures from a home environment.

Rather than creating better questionnaire batteries, other highly experimental autism research has attempted to more directly mimic the diagnostic pattern recognition employed by a human clinician by using machine learning techniques on data generated by digital sensors, audio, and video for reviews see Fusaroli, Lambrechts, Bang, Bowler, & Gaigg, 2017; Hyde et al., 2019; Jaliaawala & Khan, 2020; Kanchanamala & Sagar, 2019; Koumpouros & Kafazis, 2019. For example, classifications of eye movements (Frazier et al., 2016, 2018; Liu, Li, & Yi, 2016), speech (Fusaroli et al., 2017; Nakai, Takiguchi, Matsui, Yamaoka, & Takada, 2017), body movements (Vabalas, Gowen, Poliakoff, Casson, 2019; Vabalas, Gowen, Poliakoff, & Casson, 2020; Zhao et al., 2019), emotional expression (Jarraya, Masmoudi, & Hammami, 2020), and social interactions (Georgescu et al., 2019) have been studied in engineering and computing fields. These attributes are often measured by established technology (e.g., actigraphy) using standard machine learning pipelines, but preliminary studies have also combined these with new technology such as virtual reality (Alcañiz Raya, Chicchi Giglioli, et al., 2020; Alcañiz Raya, Marín-Morales, et al., 2020; Raya et al., 2020; Yang et al., 2017). Deep learning techniques have also been used with raw video files of children to classify behaviors (Cook, Mandal, Berry, & Johnson, 2019; Li et al., 2020; Preetham, George, George, & Verma, 2017).

While behavioral pattern identification is clearly a clinically useful target for diagnosis, proportionately more research overall has been directed towards the study of MRI where machine learning techniques have been employed widely in order to discover atypical patterns (e.g., autism spectrum and ADHD) with the ultimate aim to translate the models into useful diagnostic tools. A benefit of MRI is that the data are standardized into formats that allow databases to be built and shared. For example, the majority of studies included in the scoping review were from the open-access database Autism Imaging Data Exchange (ABIDE) database (for other reviews see Moon, Hwang, Kana, Torous, & Kim, 2019; Pagnozzi, Conti, Calderoni, Fripp, & Rose, 2018; Wolfers et al., 2019) or the ADHD-200 public database (Brown et al., 2012).

Studies suggest that MRI may be useful for early recognition, diagnosis, and differential diagnoses. For autism spectrum disorder, classification accuracies are commonly reported at 70% or above since the earliest studies (Ecker, Marquand, et al., 2010; Ecker, Rocha-Rego, et al., 2010; Jiao et al., 2010; Moon et al., 2019; Razi, Othman, & Wahab, 2015). In ADHD, a recent review that also included EEG measurements (Pulini, Kerr, Loo, & Lenartowicz, 2019) reported classification accuracies in the range of 60%–80%, but noted decreased accuracy in larger samples suggestive of experimental bias. Recent research has also employed deep learning techniques (Riaz, Asad, Alonso, & Slabaugh, 2018, 2020) with interesting studies have used data augmentation methods to overcome sample size limitations (Cicek, Ozmen, & Akan, 2019; Zhu & Chang, 2019). Early autism spectrum diagnoses from MRI in infancy has also been suggested in preliminary research (Jin, Wee, Shi, Thung, Ni, et al., 2015; Jin, Wee, Shi, Thung, Yap, et al., 2015; Shen et al., 2017, 2018) and differential diagnostic classifiers that display clinical utility by demonstrating specificity have also been developed for autism spectrum (Kushki et al., 2019; Rabany et al., 2019; Sutoko et al., 2019; Yassin et al., 2020) and ADHD (Diler et al., 2007; Duda et al., 2017; Duda, Ma, Haber, & Wall, 2016b; Faedda et al., 2016; Studerus et al., 2018).

A limitation of imaging findings is that external validation and the broader assessment of generalizability (e.g., cultural or gender differences) has not been as extensive as for questionnaire measures in autism. Exceptions to this are examples of multisite assessment procedures have been used in autism (Bhaumik, Pradhan, Das, & Bhaumik, 2018; Niu et al., 2020; Wang, Xiao, & Wu, 2019) and there are a small number of studies that have used external validation samples either testing on different ABIDE releases or on in-house datasets (Alvarez-Jimenez, Múnera-Garzón, Zuluaga, Velasco, & Romero, 2020; Bernas, Aldenkamp, & Zinger, 2018; Jahedi, Nasamran, Faires, Fan, & Müller, 2017; Plitt, Barnes, & Martin, 2015; Sadeghi et al., 2017; Shen et al., 2018). In ADHD, external validation has been conducted in a minority of studies (Cai, Chen, Szegletes, Supekar, & Menon, 2015; Yoo, Kim, Kim, & Jeong, 2019). Generalizability has also been further assessed in terms of gender (Calderoni

et al., 2012; Chaddad, Desrosiers, & Toews, 2017) and intelligence differences (Calderoni et al., 2012). These studies demonstrate a promising ability for neuroimaging to follow the lead of questionnaire measures in being clinically translated.

Prognosis

Examples of prognostic predictive machine learning in child and adolescent psychiatry can be found in the early psychosis and depression fields. Psychosis is investigated as part of early intervention initiatives (Correll et al., 2018; McGorry & Mei, 2018; McGorry et al., 2018) that aim to identify individuals at clinical high risk of a psychotic episode, mainly defined by delusions and hallucinations. There is currently no clinical method to identify transition cases and doing so has the possibility of ameliorating, or potentially preventing, the full transition to a severe illness through therapy and behavioral intervention (McGorry & Mei, 2018). For depression, the main prognostic target is suicidal symptoms and behaviors. Providing accurate and generalizable tools to predict these outcomes in individuals would have major implications to practice that could impact individuals' lives.

Machine learning for the clinical high-risk state has generated substantial prognostic research, as reviewed recent meta-analyses (Sanfelici, Dwyer, Antonucci, & Koutsouleris, 2020). Prognostic studies thus far (i.e., transition to psychosis or poor functioning) have used questionnaire measures (Koutsouleris et al., 2018, 2021), MRI (Gothelf et al., 2011; Koutsouleris et al., 2015, 2018), EEG (Ramyead et al., 2016), and biological measures such as lipids (Amminger et al., 2015). The use of the array of audio, video, and sensors is less prominent in the psychosis field, but notable innovative research has used Facebook messages to predict relapses (Birnbaum et al., 2019). In a recent meta-analysis of the psychosis high-risk area (Sanfelici et al., 2020), an average accuracy of 73% was demonstrated for prognostic predictions and sensitivity was noted to be 10% higher than using traditional statistical techniques (i.e., Cox regression). While additional studies assessing model generalizability, biases, and external validation are required, the psychosis prediction field is thus primed for robust external validation studies and prospective trials that are currently underway in order to facilitate clinical translation.

Prediction of depressive symptoms has a long history in the field with cross-validated neural network studies dating back to 1994 from questionnaire data (Kashani, Nair, Rao, Nair, & Reid, 1996; Wong & Whitaker, 1994). Early illness detection machine learning studies with large samples have also been conducted using questionnaires (McKenzie et al., 2011) and MRI using smaller samples for the prediction of future symptoms (Bertocci et al., 2016; Foland-Ross et al., 2015; Koutsouleris et al., 2018). However, the main focus in depression has been specifically on suicidality due to a strong clinical need. Suicide predictions have been conducted mainly using questionnaires (Hardt, Herke, & Schier, 2011) often in large samples (>30,000) (Walsh, Ribeiro, & Franklin, 2018), in subgroups such as medical students (Marcon et al., 2020) and minority groups (Smith, Wang, Carter, Fox, & Hooley, 2020), and when using robust nested cross-validation schemes (Miche et al., 2020). Other work has used speech processing methods (see Box 3) to identify self-injurious text (Franz, Nook, Mair, & Nock, 2020), demonstrating promisingly high accuracies in predicting suicidality (e.g., area under the curve of >0.80; Miche et al., 2020). However, as outlined in a recent meta-analysis across age-groups (Belsher et al., 2019), the accuracies belie a positive predictive value that was on average 0.01 (i.e., a 1% chance that a positive prediction will result in suicide) and this challenges the value of introducing such tools into clinical care.

Treatment prediction

In the context of treatment outcome variability with pharmaceutical and psychotherapeutic options, better matching optimal treatments to patients is important and could avoid trial-and-error strategies (Chekroud et al., 2021). Despite this, machine learning studies are relatively limited across age-groups—for a recent review see Chekroud et al. (2021). In child and adolescent psychiatry, this research gap is particularly pronounced, but with some notable exceptions. In ADHD, initial studies have been conducted to predict methylphenidate symptom remission using clinical and demographic data (Wong et al., 2017) and sleep side-effects (Yoo et al., 2020) using multiple data types (cognition, genetics, and neuroimaging) with high accuracies (>80%). In other conditions, interesting analyses of EHRs for the prediction of treatment failure in a sample of 638 children with early onset psychosis has also been conducted demonstrating the possibility of using measures that are automatically collected as part of normal clinical routine (Downs et al., 2019). For autism, new machine learning protocols have also been trialed during robot-assisted therapies (Di Nuovo, Conti, Trubia, Buono, & di Nuovo, 2018; Rudovic, Zhang, Schuller, & Picard, 2019) and using augmented reality devices paired with machine learning enhancements (Voss et al., 2019).

Summary and limitations

Reviewing the field of translational machine learning in child and adolescent psychiatry reveals increasing attempts to assist with diagnoses, prognoses, and treatment selection with new approaches and data

sources. While traditional data types are used most widely, such as questionnaires and neuroimaging, the field demonstrates emerging use of a variety of increasingly available data types that harness multilevel clinical information (e.g., data from sensors now commonly available on smartphones, video, and audio). A tool for autism diagnoses has already been FDA approved, which joins the growing array of approvals for artificial intelligence tools across various fields in medicine (Topol, 2019). When combined with the exponential rise in publications in the field more broadly, such regulatory applications suggest that implementation might be closer than once thought. The prospect of these tools being clinically used highlights the need to understand the methods used to produce the models as described above, but also to acknowledge limitations.

Sample size and representativeness

Sample size cannot be assumed to match the requirements set out in traditional statistical prediction approaches related to events-per-variable (Ogundimu et al., 2016), which often necessitate feature set reduction using a priori hypothesis-driven approaches (Fusar-Poli et al., 2019). These approaches were designed for simple statistical models, such as a Cox regression (Ogundimu et al., 2016), rather considering their regularized forms discussed above or most other machine learning algorithms. In machine learning, defining optimal sample sizes for later generalizability is still an open question (Fusar-Poli et al., 2018; Poldrack et al., 2020) that will ultimately be linked to the specific methods used and will most likely be evaluated empirically using simulation approaches.

In the absence of simple rules-of-thumb from inferential statistics, internal validation is conducted in order to avoid overfitting induced by optimized model pipelines and to empirically assess whether a model will generalize to new cases. However, the procedure is only as good as the representativeness of the sample that it was used in. For example, if the sample originates from a single site (e.g., one hospital), or contains a homogeneous subgroup of carefully selected cases, then the possibility of overfitting to the characteristics of this sample are high. Meta-analyses and simulation studies have consistently demonstrated that circumstances with small (e.g., <200 cases; Poldrack et al., 2020) or unrepresentative samples result in inflated predictive accuracy despite cross-validation (Kambeitz et al., 2015; Sanfelici et al., 2020; Schnack & Kahn, 2016; Varoquaux, 2018) and also overfitting (Cawley & Talbot, 2010). This is especially important for deep learning (Box 2) where optimal cross-validation schemes (e.g., k-fold) sometimes cannot be conducted due to computer processing demands. Within this review, inflated internal validation estimates were seen across highly experimental, pilot studies using new data types (e.g., raw video) that report unrealistically high accuracies (e.g., >90% and up to 100% accuracy) and are also most likely present when using established data types such as questionnaires (Sanfelici et al., 2020) and neuroimaging (Pulini et al., 2019).

When reading the machine learning literature, researchers need to be aware of the limitations of unrepresentative samples, which are commonly also of a small size (n < 200), but ultimately the answer to the number of individuals required depends on the rationale, aims, methods, data, and conclusions of the study. If the study is experimental (e.g., human-robot interactions; Rudovic et al., 2019), or claims of generalizability are minimal, then the number of subjects can be limited to display proof-of-concept results that could be used to design a larger study. Discouraging this research would also prevent innovation in the field towards solutions that might be ultimately the most promising. Whereas, if there are more extended claims of generalizability or clinical utility then it is necessary to note meta-analytic relationships with sample size in the field and more carefully consider internal validation, external validation, biases, and generalizability across different contexts.

Internal and external validation considerations

Internal validation using cross-validation is the standard used for preventing overfitting in machine learning. However, the robustness and quality of internal cross-validation needs to be considered in order to assess the possibility of external validation and ultimately translation. The most common mistake in internal cross-validation within psychiatry is when features are chosen on the basis of the target variable using traditional methods in the sample (e.g., t-tests) and then they are used separately in a cross-validation procedure for predictive purposes—for example, choosing brain regions or questionnaire items on the basis of initial pair-wise comparisons and then forwarding only these variables to a cross-validated machine learning pipeline. This 'double-dipping' is an example of severe information leakage between the training and test samples, which results in invalid and overfitted results. Widespread circularity in conclusions due to this problem and others has been reported in the child and adolescent neuroimaging field (Pulini et al., 2019), and is likely across other data domains, thus requiring caution when assessing results that have only been internally validated. For this reason and others, some journals (e.g., Lancet Psychiatry) will not accept results that only contain internal validation approaches and require external validation.

A second consideration is to assess the internal cross-validation procedure itself. Research has demonstrated that leave-one-case-out cross-validation results in inflated estimates of the true

generalizability to new cases (Varoquaux et al., 2017). While k-fold cross-validation can reduce this possibility, there is a need for designs that can provide uncertainty measurements for individual predictions in order to determine stability (e.g., standard deviation of accuracy estimates). For this reason, there is a need to apply such schemes as repeated (and preferably nested) designs in order to more thoroughly test the predictive capacity and provide accurate central tendency measures (i.e., mean or median) in addition to measures of variance. Studies using leave-one-site-out designs within a repeated, nested cross-validation are an example of where study-specific internal cross-validation can provide enhanced estimates of generalizability because they simulate the process of applying a model within a new context. Validation hierarchies can be used to assess studies in this regard (Dwyer et al., 2018).

Despite the cross-validation design, the gold-standard is still external validation in an independent study where the same models are applied to new individuals. Within this context, it still remains important to assess the degree of generalizability across samples that the results will deliver. External validation in another highly similar, unrepresentative sample that is very close in geographic or cultural proximity to the discovery sample obviously limits translational claims that go beyond these samples to the general population. Whereas, if the external validation sample is from a different cultural context that contains multiple sites (e.g., in a consortium study), then the level of generalizability can be judged higher. Ultimately, as demonstrated from this review, multisite prospective clinical trials will most likely be required for regulatory approval (e.g., the Cognoa diagnostic tool for autism; www.cognoa.com).

Clinical utility

The scoping review demonstrated a publication bias towards brain measures measured with MRI and EEG. While these techniques are used in clinical practice and could be part of the future of translational machine learning (Walter et al., 2019), there are implementation challenges, such as cost, clinical access, and patient burden. Since there are less burdensome alternatives (e.g., questionnaires), combined approaches may be most beneficial (Koutsouleris et al., 2021). A good example highlighting this point is in the first machine learning competition in the field of neuroimaging that aimed to classify individuals with ADHD. While multiple strategies produced promising classification accuracies, the winning strategy overall simply used demographic details and intelligence measures (Brown et al., 2012). In reflection of a real-life clinical workflow that only conducts burdensome new tests if they are clinically indicated (e.g., a brain scan if a tumor is suspected), these findings highlight the potential importance of sequential clinical pipelines that only suggest a new test if it is statistically indicated for each individual (Koutsouleris et al., 2021). Future directions in the area of clinical utility could also involve assessments that are easy to conduct and occur in the home environment (Abbas et al., 2020) and are specifically tailored to children and adolescents (e.g., with video game play; Aggarwal, Saluja, Gambhir, Gupta, & Satia, 2020). Net-benefit analyses would also be beneficial (Fusar-Poli et al., 2018) in addition to more comprehensive assessments of implementation challenges (Beede et al., 2020).

Ethical concerns

As detailed in other reviews (Cath, 2018; Cohen, Amarasingham, Shah, Xie, & Lo, 2014; Price & Cohen, 2019), there are major ethical concerns with model translation that need to be considered. A central issue is related to bias in translational science against specific ethnic, cultural, or gender groups due to a lack of diversity included in training samples or appropriate assessments of algorithm bias (Cahan, Hernandez-Boussard, Thadaney-Israni, & Rubin, 2019). If the machine learns from the majority groups of a population then it will potentially make mistakes with minority groups, which should be investigated and then mitigated or transparently described before model deployment. In medicine, machine learning biases have been demonstrated for ethnicity in critical studies highlighting negative real-world consequences (Obermeyer, Powers, Vogeli, & Mullainathan, 2019) and also for gender (Cirillo et al., 2020). Within this context, however, it is important to also note that biases need to be considered for any predictive tool or study, as demonstrated by ethically questionable biases in the Framingham risk score (Gijsberts et al., 2015), in precision genomics (Martin et al., 2019), and in neuroimaging (Crossley et al., 2019). As such, it is essential that biases are considered carefully in any science that claims to have translational potential and especially in highly translational fields, such as machine learning. A second major issue concerns prognoses and whether it is ethical to provide predictions given the possible iatrogenic effects of the prediction especially for psychiatric conditions (Martinez-Martin, Dunn, & Roberts, 2018). When combined, continued investment in ethical oversight and governance needs to be considered for psychiatry as much as for other fields of medicine where artificial intelligence solutions are closer to widespread deployment.

Future directions for prediction

The question of whether machine learning will make changes to medical care has been around since the beginnings of the techniques (Shortliffe, 1993) and it

remains an open question now. However, this review has demonstrated that this future may be closer than once thought (Topol, 2019). To facilitate the continued success of the approach, future research could involve addressing the limitations above to facilitate collaborative research and to invest in machine learning research directions with specific relevance to child and adolescent psychiatry.

Programs and platforms

Standard statistical tools that do not involve computer programming skills (e.g., SPSS) cannot be used for many machine learning tasks and most research to-date has been produced using programming languages. Multiple toolboxes exist for the major programming languages used by researchers, such as scikit-learn for Python (https://scikit-learn.org/), caret for R (https://cran.r-project.org/web/packages/caret/index.html), or the Machine Learning Package for MATLAB (https://www.mathworks.com/solutions/machine-learning.html), in addition to deep-learning specific tools such as Keras (https://keras.io/). These tools make it easier to flexibly and creatively implement machine learning approaches, in addition to providing a community of similar users and facilitate code transparency. However, the limitation of the packages for new researchers to the field, or those with little coding experience, is that the flexibility of the code can lead to errors in pipeline development and can hinder model sharing with researchers who have no coding experience—that is, to enable external validation in another sample.

To facilitate machine learning analyses that do not require coding experience and provide standardized pipelines, there are tools that have been specifically developed for the psychiatric field, such as PRoNTO (Schrouff et al., 2013) and NeuroMiner (https://github.com/neurominer-git). These tools provide graphical user interfaces that allow users to enter data and design an analysis (i.e., in similarity to SPSS). In addition, a deep learning tool developed by psychiatric researchers that reduces the burden of programming is PHOTON (https://photon-ai.com/). The benefit of using psychiatric machine learning software is that they are better tailored to some of the main questions encountered; but in addition, they increase the transparency and reproducibility of analyses because the basic pipelines and operations (e.g., establishing cross-validation pipelines) have been established.

Once algorithms have been created using programming languages or software with graphical user interfaces, they need to be made available to other researchers in order to apply them to their own data. In the case of machine learning models, this is sometimes challenging because there can be thousands of models from cross-validation schemes representing analysis pipelines that convert dense data to meaningful predictions (He et al., 2019). By combining the software tools with online platforms for model sharing and application (e.g., see www.proniapredictors.eu or https://photon-ai.com/), future research would benefit from allowing widespread generalizability testing required for clinical application and ultimately clinical deployment.

Data aggregation and federated analyses

Future machine learning analysis will increasingly rely on data sharing through aggregation and federated analyses. Firstly, data sharing needs to increase in order to enhance sample sizes, representativeness of the population, and generalizability testing using both internal and external validation. For example, aggregated databases outlined in this review (e.g., ADHD-200 or ABIDE) have enabled a community of analysts from across disciplines to contribute a large amount of research and test the generalizability of models across study sites. Specifically, 25% (n = 22) of all neuroimaging studies in autism were from the ABIDE cohort and 58% (n = 35) of ADHD imaging studies were from the ADHD-200 repository that has continued to grow with additional sites since it was first released (Brown et al., 2012). In cases where the data cannot be shared, funded initiatives to facilitate realistic collaboration between studies and consortia will be required—ideally across different countries (e.g., the SCZ-AMP NIMH initiative)(Woods, Choi, & Mamah, 2021). These efforts could involve centralized data aggregation across modalities or alternatively could leverage existing software to conduct decentralized federated analyses where data from separate studies is stored locally and models are built either in a cloud (e.g., ViPAR; Carter et al., 2015) or they are built locally and then only the model parameters are combined (e.g., DataShield; Wolfson et al., 2010). To support these efforts, further investment needs to occur in software development to enhance the pre-existing solutions and build towards more widespread adoption.

Transdiagnostic and interdisciplinary analyses

A notable aspect of the review was the recognition that cross-talk between disciplines would be highly recommended. For example, although an ADHD diagnosis is partially based on behavioral criteria questionnaire reduction techniques, wearables, and video or movement assessments were not found to be widely studied and could be imported from autism research. Prognostic predictions are also limited in both ADHD and autism fields despite a great need, thus methods could be taken from the psychosis and depression literature where multimodal assessment techniques are used to predict outcomes. Similarly, the prediction of symptoms is critical across fields, but is most often investigated in depression and these methods could be transferred. A substantially

missing area generally across domains was in the field of treatment selection and outcomes, which requires further work given the vital importance of providing more personalized treatment recommendations. All fields can also learn from the productivity of releasing large public databases and encouraging their use through competitions (e.g., ADHD-200 and ABIDE).

Multimodal analyses and diagnostic chains

An exciting vision of a future machine learning procedure could involve multiple sequential steps that implement cost-effective sequences of clinical and biological assessments to maximize predictive power and clinical utility (Abbas et al., 2020; Koutsouleris et al., 2021). There are a small number of child and adolescent studies that have enhanced MRI predictions with additional modalities such as genetic data (Yoo et al., 2019), questionnaires and cognition (Farzi, Kianian, & Rastkhadive, 2017), and genetics, questionnaires, and cognition (Yoo et al., 2020). Ultimately, it is notable that the only tool in child and adolescent psychiatry that has been approved by the FDA involves the aggregation of data from multiple assessments (Abbas et al., 2020). Further research in this area could help to move the field further towards translation.

Normative modeling, unsupervised learning, and transfer learning

In terms of machine learning methods, one notable area that could receive more attention specifically for child and adolescent psychiatry is the field of normative modeling (Marquand, Rezek, Buitelaar, & Beckmann, 2016). This machine learning technique maps deviations from normal development in order to characterize abnormality in much the same way as a growth chart. As such, it is possible to chart brain age, for example, and then determine how an individual differs from the normative age trajectory (i.e., to estimate 'brain age' in addition to chronological age). Similarly, unsupervised learning research is currently limited despite it having success in other medical fields to identify subgroups of adults based on brain patterns (Chand et al., 2020), clinical data (Dwyer et al., 2020), and multimodal data (Luo et al., 2020). Further research needs to be conducted, for example, to build on research in children that has been conducted to identify behavioral phenotypes of autism in order to address questions related to diagnostic heterogeneity (Stevens et al., 2017, 2019). A critical effort also needs to be made in investigating whether models from adults could be used in children and adolescents in order to leverage model development across the lifespan (this is known as transfer learning; Pan & Yang, 2009). Transfer learning would also be recommended across diagnoses in order to determine whether the field could capitalize on combining datasets for specific problems (e.g., prognoses).

Conclusions

Translational machine learning for psychiatry is a paradigm that attempts to turn data into actionable clinical information using computers. The algorithmic approaches are designed to produce optimally generalizable predictions from data that range in structure from hypothesis-driven feature sets based on questionnaire measures to unstructured EHR notes or biological data. In child and adolescent psychiatry, the techniques are increasingly being used for diagnoses, prognoses, and treatment selection purposes, with one tool having received FDA-approval. Pilot research also gives a perspective into the future of practice through the use of data from video, audio, virtual reality, game play, and even human-robot interactions. Caution regarding the translational potential is required until gold-standard levels of validation are achieved and ethical issues are addressed, which could be facilitated through further collaboration between groups of researchers who aim to use the statistical approaches for prediction to directly assist in clinical care (Box 4). Despite the specific techniques employed, the paradigm is likely to have a lasting impact on a field that is only beginning to turn

Box 4 Revisiting the case example

Picture a clinician seeing the same adolescent from the Box 1 case example who may have been at risk for psychosis. Instead of turning to the literature or guidelines to understand the group-based risk profile, the clinician is aware of algorithmic tools that can be used as decision aides. They then conduct a gender-specific screening test on a tablet involving a small number of questions. Prior to the assessment the adolescent and their family has also completed short questionnaires on their smartphones. During the clinician's regular clinical assessment, the data is automatically combined and analyzed using machine learning algorithms in real-time to produce an intuitive report that quantifies the adolescent's risk for psychosis with margins of uncertainty, suggests further assessments to increase the certainty of predictions, and indicates successful therapeutic options based on their specific profile. The clinician then integrates this information with their broader assessment, incorporates it into their clinical report, and is ultimately able to develop a targeted clinical plan that is tailored to the delay and prevention of a possible transition to psychosis.

statistical associations into actionable clinical decisions for individual patients.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Search terms for the scoping review.

Acknowledgements

This work was supported by United States National Institutes of Health grant R01MH112070. The work was also supported by the PRONIA project as funded by the European Union 7th Framework Program grant 602152. The authors would like to acknowledge Elif Sarisik for her helpful comments on the document. The authors have declared that they have no competing or potential conflicts of interest. DD is supported by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (#30196). NK is supported through grants from NIH (U01MH124639-01; ProNET), the Wellcome Trust, the German Innovation Fund (CARE project), the German Federal Ministry of Education and Research (COMMITMENT and BEST projects), as well as ERA PerMed (IMPLEMENT project).

Correspondence

Dominic Dwyer, Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Nussbaumstr. 7, D-80336 Munich. Email: domdwyer@gmail.com

Nikolaos Koutsouleris, Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Nussbaumstr. 7, D-80336 Munich, Germany. Email: Nikolaos.koutsouleris@med.uni-muenchen.de

Key points

- Providing personalized recommendations for children and adolescents could be critical to their development.
- Traditional statistical techniques have not been translated clinically to facilitate changes in clinical care.
- Machine learning approaches may help to provide individualized recommendations for diagnoses, prognoses, and treatments.
- The approaches are characterized by flexible algorithm pipelines that address limitations of traditional statistical approaches and aim to provide the most accurate and generalizable fit to the data.
- Machine learning research is currently exponentially rising in child and adolescent psychiatry using a wide range of data types.
- Specific studies demonstrate the potential of the techniques to lead to translation, with one machine learning tool being approved by national medical regulators to be used for autism.
- Limitations of the techniques need to be considered and future directions would benefit from collaborative research endeavors.

References

Abbas, H., Garberson, F., Glover, E., & Wall, D.P. (2017). Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening. In J.Y. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baezayates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, & M. Toyoda (Eds.) *2017 IEEE International Conference on Big Data.* (pp. 3558–3561).

Abbas, H., Garberson, F., Glover, E., & Wall, D.P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, *25*, 1000–1007.

Abbas, H., Garberson, F., Liu-Mayo, S., Glover, E., & Wall, D.P. (2020). Multi-modular AI approach to streamline autism diagnosis in young children. *Scientific Reports*, *10*, 5014.

Abi-Dargham, A., & Horga, G. (2016). The search for imaging biomarkers in psychiatric disorders. *Nature Medicine*, *22*, 1248–1255.

Achenie, L.E.K., Scarpa, A., Factor, R.S., Wang, T., Robins, D.L., & McCrickard, D.S. (2019). A machine learning strategy for autism screening in toddlers. *Journal of Developmental and Behavioral Pediatrics*, *40*, 369–376.

Aggarwal, S., Saluja, S., Gambhir, V., Gupta, S., & Satia, S.P.S. (2020). Predicting likelihood of psychological disorders in PlayerUnknown's Battlegrounds (PUBG) players from Asian countries using supervised machine learning. *Addictive Behaviors*, *101*, 106132.

Alcañiz Raya, M., Chicchi Giglioli, I.A., Marín-Morales, J., Higuera-Trujillo, J.L., Olmos, E., Minissi, M.E., … & Abad, L. (2020). Application of supervised machine learning for behavioral biomarkers of autism spectrum disorder based on electrodermal activity and virtual reality. *Frontiers in Human Neuroscience*, *14*, 90.

Alcañiz Raya, M., Marín-Morales, J., Minissi, M.E., Teruel Garcia, G., Abad, L., & Chicchi Giglioli, I.A. (2020). Machine learning and virtual reality on body movements' behaviors to classify children with autism spectrum disorder. *Journal of Clinical Medicine*, *9*, 1260.

Alvarez-Jimenez, C., Múnera-Garzón, N., Zuluaga, M.A., Velasco, N.F., & Romero, E. (2020). Autism spectrum disorder characterization in children by capturing local-regional brain changes in MRI. *Medical Physics*, *47*, 119–131.

Amminger, G.P., Mechelli, A., Rice, S., Kim, S.W., Klier, C.M., McNamara, R.K., … & Schafer, M.R. (2015). Predictors of treatment response in young people at ultra-high risk for psychosis who received long-chain omega-3 fatty acids. *Translational Psychiatry*, *5*, e495.

Barness, L.A., Tunnessen, W.W., Jr, Worley, W.E., Simmons, T.L., & Ringe, T.B., Jr (1974). Computer-assisted diagnosis

in pediatrics. *American Journal of Diseases of Children*, *127*, 852–858.
Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L.M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. (pp. 1–12).
Belsher, B.E., Smolenski, D.J., Pruitt, L.D., Bush, N.E., Beech, E.H., Workman, D.E., … & Skopp, N.A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, *76*, 642–651.
Ben-Sasson, A., Robins, D.L., & Yom-Tov, E. (2018). Risk assessment for parents who suspect their child has autism spectrum disorder: Machine learning approach. *Journal of Medical Internet Research*, *20*, e134.
Bernas, A., Aldenkamp, A.P., & Zinger, S. (2018). Wavelet coherence-based classifier: A resting-state functional MRI study on neurodynamics in adolescents with high-functioning autism. *Computer Methods and Programs in Biomedicine*, *154*, 143–151.
Bertocci, M.A., Bebko, G., Versace, A., Fournier, J.C., Iyengar, S., Olino, T., … & Phillips, M.L. (2016). Predicting clinical outcome from reward circuitry function and white matter structure in behaviorally and emotionally dysregulated youth. *Molecular Psychiatry*, *21*, 1194–1201.
Bhaumik, R., Pradhan, A., Das, S., & Bhaumik, D.K. (2018). Predicting autism spectrum disorder using domain-adaptive cross-site evaluation. *Neuroinformatics*, *16*, 197–205.
Bird, J.C., Fergusson, E.C., Kirkham, M., Shearn, C., Teale, A.L., Carr, L., … & Freeman, D. (2021). Paranoia in patients attending child and adolescent mental health services. *Australian and New Zealand Journal of Psychiatry*, 4867420981416. https://doi.org/10.1177/0004867420981416
Birnbaum, M.L., Ernala, S.K., Rizvi, A.F., Arenare, E., R. Van Meter, A., De Choudhury, M., & Kane, J.M. (2019). Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. *NPJ Schizophrenia*, *5*, 17.
Bone, D., Bishop, S.L., Black, M.P., Goodwin, M.S., Lord, C., & Narayanan, S.S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, *57*, 927–937.
Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. (pp. 144–152). ACM.
Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199–215.
Brown, M.R., Sidhu, G.S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P.H., … & Dursun, S.M. (2012). ADHD-200 Global Competition: Diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Frontiers in Systems Neuroscience*, *6*, 69.
Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*, 233–234.
Bzdok, D., Engemann, D., & Thirion, B. (2020). Inference and prediction diverge in biomedicine. *Patterns*, *1*, 100119.
Bzdok, D., & Ioannidis, J.P. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in Neurosciences*, *42*, 251–262.
Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*, 223–230.
Cahan, E.M., Hernandez-Boussard, T., Thadaney-Israni, S., & Rubin, D.L. (2019). Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digital Medicine*, *2*, 78.
Cai, W., Chen, T., Szegletes, L., Supekar, K., & Menon, V. (2015). Aberrant cross-brain network interaction in children with attention-deficit/hyperactivity disorder and its relation to attention deficits: a multisite and cross-site replication study. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2015.10.017
Calderoni, S., Retico, A., Biagi, L., Tancredi, R., Muratori, F., & Tosetti, M. (2012). Female children with autism spectrum disorder: an insight from mass-univariate and pattern classification analyses. *NeuroImage*, *59*, 1013–1022.
Carter, K.W., Francis, R.W., Carter, K.W., Francis, R.W., Bresnahan, M., Gissler, M., … & Yusof, Z. (2015). ViPAR: A software platform for the Virtual Pooling and Analysis of Research Data. *International Journal of Epidemiology*, *45*, 408–416.
Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*, 20180080.
Cawley, G.C., & Talbot, N.L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.
Chaddad, A., Desrosiers, C., & Toews, M. (2017). Multi-scale radiomic analysis of sub-cortical regions in MRI related to autism, gender and age. *Scientific Reports*, *7*, 45639.
Chand, G.B., Dwyer, D.B., Erus, G., Sotiras, A., Varol, E., Srinivasan, D., … & Davatzikos, C. (2020). Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. *Brain*, *143*, 1027–1038.
Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*, 16–28.
Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., … & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*, 154–170.
Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., … & Corlett, P.R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *Lancet Psychiatry*, *3*, 243–250.
Cicek, G., Ozmen, A., Akan, A., & IEEE (2019). The effect of data augmentation on ADHD diagnostic model using deep learning.
Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., … & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine*, *3*, 81.
Cohen, I.G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, *33*, 1139–1147.
Cohen, I.L., Sudhalter, V., Landon-Jimenez, D., & Keogh, M. (1993). A neural network approach to the classification of autism. *Journal of Autism and Developmental Disorders*, *23*, 443–466.
Cook, A., Mandal, B., Berry, D., & Johnson, M. (2019). Towards automatic screening of typical and atypical behaviors in children with autism. In L. Singh, R. Deveaux, G. Karypis, F. Bonchi, & J. Hill (Eds.) *2019 IEEE International Conference on Data Science and Advanced Analytics*. (pp. 504–510).
Corcoran, C.M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., … & Cecchi, G.A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*, 67–75.
Correll, C.U., Galling, B., Pawar, A., Krivko, A., Bonetto, C., Ruggeri, M., … & Kane, J.M. (2018). Comparison of early intervention services vs treatment as usual for early-phase psychosis: a systematic review, meta-analysis, and meta-regression. *JAMA Psychiatry*, *75*, 555–565.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cox, G.R., Callahan, P., Churchill, R., Hunot, V., Merry, S.N., Parker, A.G., & Hetrick, S.E. (2014). Psychological therapies versus antidepressant medication, alone and in combination for depression in children and adolescents. *Cochrane Database of Systematic Reviews*, *2014*, CD008324.

Crossley, N.A., Alliende, L.M., Ossandon, T., Castañeda, C.P., González-Valderrama, A., Undurraga, J., … & Bressan, R. (2019). Imaging social and environmental factors as modulators of brain dysfunction: Time to focus on developing non-western societies. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*, 8–15.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Dalsgaard, S., Thorsteinsson, E., Trabjerg, B.B., Schullehner, J., Plana-Ripoll, O., Brikell, I., … & Pedersen, C.B. (2020). Incidence rates and cumulative incidences of the full spectrum of diagnosed mental disorders in childhood and adolescence. *JAMA Psychiatry*, *77*, 155–164.

Dawson, G., & Bernier, R. (2013). A quarter century of progress on the early detection and treatment of autism spectrum disorder. *Development and Psychopathology*, *25*, 1455–1472.

di Nuovo, A., Conti, D., Trubia, G., Buono, S., & di Nuovo, S. (2018). Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability. *Robotics*, *7*, 25.

Diler, R.S., Daviss, W.B., Lopez, A., Axelson, D., Iyengar, S., & Birmaher, B. (2007). Differentiating major depressive disorder in youths with attention deficit hyperactivity disorder. *Journal of Affective Disorders*, *102*, 125–130.

Downs, J., Dean, H., Lechler, S., Sears, N., Patel, R., Shetty, H., … & Pina-Camacho, L. (2019). Negative symptoms in early-onset psychosis and their association with antipsychotic treatment failure. *Schizophrenia Bulletin*, *45*, 69–79.

Duda, M., Haber, N., Daniels, J., & Wall, D.P. (2017). Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Translational Psychiatry*, *7*, e1133.

Duda, M., Kosmicki, J.A., & Wall, D.P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational Psychiatry*, *4*, e424.

Duda, M., Ma, R., Haber, N., & Wall, D.P. (2016a). Use of machine learning for behavioral distinction of autism and ADHD. *Translational Psychiatry*, *6*, e732.

Duda, M., Ma, R., Haber, N., & Wall, D.P. (2016b). Use of machine learning for behavioral distinction of autism and ADHD. *Translational Psychiatry*, *6*, e732.

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

Dwyer, D.B., Kalman, J.L., Budde, M., Kambeitz, J., Ruef, A., Antonucci, L.A., … & Koutsouleris, N. (2020). An investigation of psychosis subgroups with prognostic validation and exploration of genetic underpinnings: The PsyCourse study. *JAMA Psychiatry*, *77*, 523–533.

Ecker, C., Marquand, A., Mourão-Miranda, J., Johnston, P., Daly, E.M., Brammer, M.J., … & Murphy, D.G. (2010). Describing the brain in autism in five dimensions–magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *Journal of Neuroscience*, *30*, 10612–10623.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E.M., … & Murphy, D.G. (2010). Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. *NeuroImage*, *49*, 44–56.

Eni, M., Dinstein, I., Ilan, M., Menashe, I., Meiri, G., & Zigel, Y. (2020). Estimating autism severity in young children from speech signals using a deep neural network. *IEEE Access*, *8*, 139489–139500.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., Depristo, M., Chou, K., … & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*, 24–29.

Faedda, G.L., Ohashi, K., Hernandez, M., McGreenery, C.E., Grant, M.C., Baroni, A., … & Teicher, M.H. (2016). Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls. *Journal of Child Psychology and Psychiatry*, *57*, 706–716.

Farzi, S., Kianian, S., Rastkhadive, I., & IEEE (2017). Diagnosis of attention deficit hyperactivity disorder using deep belief network based on greedy approach.

Filzmoser, P., Liebmann, B., & Varmuza, K. (2009). Repeated double cross validation. *Journal of Chemometrics*, *23*, 160–171.

Fisher, A.J., Medaglia, J.D., & Jeronimus, B.F. (2018). Lack of group-to-individual generalizability is a threat to human Subjects research. *Proceedings of the National Academy of Sciences, USA*, *115*, E6106–e6115.

Foland-Ross, L.C., Sacchet, M.D., Prasad, G., Gilbert, B., Thompson, P.M., & Gotlib, I.H. (2015). Cortical thickness predicts the first onset of major depression in adolescence. *International Journal of Developmental Neuroscience*, *46*, 125–131.

Franz, P.J., Nook, E.C., Mair, P., & Nock, M.K. (2020). Using topic modeling to detect and describe self-injurious and related content on a large-scale digital platform. *Suicide and Life-Threatening Behavior*, *50*, 5–18.

Frazier, T.W., Klingemier, E.W., Beukemann, M., Speer, L., Markowitz, L., Parikh, S., … & Strauss, M.S. (2016). Development of an objective autism risk index using remote eye tracking. *Journal of the American Academy of Child and Adolescent Psychiatry*, *55*, 301–309.

Frazier, T.W., Klingemier, E.W., Parikh, S., Speer, L., Strauss, M.S., Eng, C., … & Youngstrom, E.A. (2018). Development and validation of objective and quantitative eye tracking-based measures of autism risk and symptom levels. *Journal of the American Academy of Child and Adolescent Psychiatry*, *57*, 858–866.

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D.M., & Gaigg, S.B. (2017). Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Research*, *10*, 384–407.

Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rossler, A., Schultze-Lutter, F., … & Yung, A. (2013). The psychosis high-risk state: A comprehensive state-of-the-art review. *JAMA Psychiatry*, *70*, 107–120.

Fusar-Poli, P., Hijazi, Z., Stahl, D., & Steyerberg, E.W. (2018). The science of prognosis in psychiatry: A review. *JAMA Psychiatry*, *75*, 1289–1297.

Fusar-Poli, P., Stringer, D., M. S. Durieux, A., Rutigliano, G., Bonoldi, I., De Micheli, A., & Stahl, D. (2019). Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk. *Translational Psychiatry*, *9*, 259.

Georgescu, A.L., Koehler, J.C., Weiske, J., Vogeley, K., Koutsouleris, N., & Falter-Wagner, C. (2019). Machine learning to study social interaction difficulties in ASD. *Frontiers in Robotics and AI*, *6*. https://doi.org/10.3389/frobt.2019.00132

Gijsberts, C.M., Groenewegen, K.A., Hoefer, I.E., Eijkemans, M.J., Asselbergs, F.W., Anderson, T.J., … & den Ruijter, H.M. (2015). Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*, *10*, e0132321.

Goodman, S. (1992). A comment on replication, P-values and evidence. *Statistics in Medicine*, *11*, 875–879.

Goodman, S.N., Fanelli, D., & Ioannidis, J.P. (2016). What does research reproducibility mean? *Science: Translational Medicine*, *8*, 341ps312.

Gothelf, D., Hoeft, F., Ueno, T., Sugiura, L., Lee, A.D., Thompson, P., & Reiss, A.L. (2011). Developmental changes

in multivariate neuroanatomical patterns that predict risk for psychosis in 22q11.2 deletion syndrome. *Journal of Psychiatric Research*, *45*, 322–331.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., … & Webster, D.R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*, 2402–2410.

Hardt, J., Herke, M., & Schier, K. (2011). Suicidal ideation, parent-child relationships, and adverse childhood experiences: A cross-validation study using a graphical markov model. *Child Psychiatry and Human Development*, *42*, 119–133.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer-Verlag New York.

He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*, 30–36.

Hollon, S.D., Areán, P.A., Craske, M.G., Crawford, K.A., Kivlahan, D.R., Magnavita, J.J., … & Kurtzman, H. (2014). Development of clinical practice guidelines. *Annual Review of Clinical Psychology*, *10*, 213–241.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., & Aerts, H.J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*, 500–510.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441.

Hyde, K.K., Novack, M.N., Lahaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D.R., & Linstead, E. (2019). Applications of supervised machine learning in autism spectrum disorder research: A review. *Review Journal of Autism and Developmental Disorders*, *6*, 128–146.

Hyman, S.E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology*, *6*, 155–179.

Insel, T.R., & Cuthbert, B.N. (2015). Brain disorders? Precisely. *Science*, *348*, 499–500.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Med*, *2*, e124.

Ioannidis, J.P.A. (2016). Why most clinical research is not useful. *PLoS Med*, *13*, e1002049.

Jahedi, A., Nasamran, C.A., Faires, B., Fan, J., & Müller, R.A. (2017). Distributed intrinsic functional connectivity patterns predict diagnostic status in large autism cohort. *Brain Connectivity*, *7*, 515–525.

Jaliaawala, M.S., & Khan, R.A. (2020). Can autism be catered with artificial intelligence-assisted intervention technology? A comprehensive survey. *Artificial Intelligence Review*, *53*, 1039–1069.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Jarraya, S.K., Masmoudi, M., & Hammami, M. (2020). Compound emotion recognition of autistic children during meltdown crisis based on deep spatio-temporal analysis of facial geometric features. *IEEE Access*, *8*, 69311–69326.

Jiao, Y., Chen, R., Ke, X., Chu, K., Lu, Z., & Herskovits, E.H. (2010). Predictive models of autism spectrum disorder based on brain regional cortical thickness. *NeuroImage*, *50*, 589–599.

Jin, Y., Wee, C.Y., Shi, F., Thung, K.H., Ni, D., Yap, P.T., & Shen, D. (2015). Identification of infants at high-risk for autism spectrum disorder using multiparameter multiscale white matter connectivity networks. *Human Brain Mapping*, *36*, 4880–4896.

Jin, Y., Wee, C.Y., Shi, F., Thung, K.H., Yap, P.T., & Shen, D. (2015). Identification of infants at risk for autism using multi-parameter hierarchical white matter connectomes. *Machine Learning and Medical Imaging*, *9352*, 170–177.

Jordan, M.I., & Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*, 255–260.

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., … & Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, *40*, 1742–1751.

Kanchanamala, P., & Sagar, G.L. (2019). A review of machine learning models for predicting autism spectrum disorder. *Helix*, *9*, 4797–4801.

Kapur, S., Phillips, A.G., & Insel, T.R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*, 1174–1179.

Kashani, J.H., Nair, S.S., Rao, V.G., Nair, J., & Reid, J.C. (1996). Relationship of personality, environmental, and DICA variables to adolescent hopelessness: A neural network sensitivity approach. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 640–645.

Kohavi, R., & John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*, 273–324.

Kosmicki, J.A., Sochat, V., Duda, M., & Wall, D.P. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry*, *5*, e514.

Koumpouros, Y., & Kafazis, T. (2019). Wearables and mobile technologies in Autism Spectrum Disorder interventions: A systematic literature review. *Research in Autism Spectrum Disorders*, *66*, 101405.

Koutsouleris, N., Dwyer, D.B., Degenhardt, F., Maj, C., Urquijo-Castro, M.F., Sanfelici, R. … & Meisenzahl, E. (2021). Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry*, *78*, 195–209.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D.B., … & Borgwardt, S. (2018). Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry*, *75*, 1156–1172.

Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E.M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., … & Borgwardt, S. (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin*, *41*, 471–482.

Kushki, A., Anagnostou, E., Hammill, C., Duez, P., Brian, J., Iaboni, A., … & Lerch, J.P. (2019). Examining overlap and homogeneity in ASD, ADHD, and OCD: A data-driven, diagnosis-agnostic approach. *Translational Psychiatry*, *9*, 318.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lee, D.D., & Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Li, J., Zhong, Y.H., Han, J.X., Ouyang, G.X., Li, X.L., & Liu, H.H. (2020). Classifying ASD children with LSTM based on raw videos. *Neurocomputing*, *390*, 226–238.

Liu, W.B., Li, M., & Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, *9*, 888–898.

Lo, A., Chernoff, H., Zheng, T., & Lo, S.H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 13892–13897.

Luo, Y., Eran, A., Palmer, N., Avillach, P., Levy-Moonshine, A., Szolovits, P., & Kohane, I.S. (2020). A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nature Medicine*, *26*, 1375–1379.

Maenner, M.J., Shaw, K.A., Baio, J., Washington, A., Patrick, M., Dirienzo, M., … & Dietz, P.M. (2020). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network,

11 sites, United States, 2016. *MMWR Surveillance Summary*, *69*, 1–12.
Maenner, M.J., Yeargin-Allsopp, M., Van Naarden Braun, K., Christensen, D.L., & Schieve, L.A. (2016). Development of a machine learning algorithm for the surveillance of autism spectrum disorder. *PLoS One*, *11*, e0168224.
Marcon, G., Monteiro, G.M.C., Ballester, P., Cassidy, R.M., Zimerman, A., Brunoni, A.R., … & Passos, I.C. (2020). Who attempts suicide among medical students? *Acta Psychiatrica Scandinavica*, *141*, 254–264.
Marín, O. (2016). Developmental timing and critical windows for the treatment of psychiatric disorders. *Nature Medicine*, *22*, 1229–1238.
Marquand, A.F., Rezek, I., Buitelaar, J., & Beckmann, C.F. (2016). Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. *Biological Psychiatry*, *80*, 552–561.
Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., & Daly, M.J. (2019). Current clinical use of polygenic scores will risk exacerbating health disparities. *Nature Genetics*, *51*, 584.
Martinez-Martin, N., Dunn, L.B., & Roberts, L.W. (2018). Is it ethical to use prognostic estimates from machine learning to treat psychosis? *AMA Journal of Ethics*, *20*, E804–811.
McGorry, P.D., & Mei, C. (2018). Ultra-high-risk paradigm: Lessons learnt and new directions. *Evidence Based Mental Health*, *21*, 131–133.
McGorry, P.D., Ratheesh, A., & O'Donoghue, B. (2018). Early intervention-an implementation challenge for 21st century mental health care. *JAMA Psychiatry*, *75*, 545–546.
McKenzie, D.P., Toumbourou, J.W., Forbes, A.B., Mackinnon, A.J., McMorris, B.J., Catalano, R.F., & Patton, G.C. (2011). Predicting future depression in adolescents using the Short Mood and Feelings Questionnaire: A two-nation study. *Journal of Affective Disorders*, *134*, 151–159.
McPartland, J.C., Reichow, B., & Volkmar, F.R. (2012). Sensitivity and specificity of proposed DSM-5 diagnostic criteria for autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*, 368–383.
Miche, M., Studerus, E., Meyer, A.H., Gloster, A.T., Beesdo-Baum, K., Wittchen, H.U., & Lieb, R. (2020). Prospective prediction of suicide attempts in community adolescents and young adults, using regression methods and machine learning. *Journal of Affective Disorders*, *265*, 570–578.
Miotto, R., Li, L., Kidd, B.A., & Dudley, J.T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*, 26094.
Molnar, C. (2020). Interpretable machine learning: Lulu.com.
Moon, S.J., Hwang, J., Kana, R., Torous, J., & Kim, J.W. (2019). Accuracy of machine learning algorithms for the diagnosis of autism spectrum disorder: systematic review and meta-analysis of brain magnetic resonance imaging studies. *JMIR Mental Health*, *6*, e14108.
Moore, M.M., Slonimsky, E., Long, A.D., Sze, R.W., & Iyer, R.S. (2019). Machine learning concepts, concerns and opportunities for a pediatric radiologist. *Pediatric Radiology*, *49*, 509–516.
Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 22071–22080.
Naci, H., & Ioannidis, J.P. (2015). How good is "evidence" from clinical studies of drug effects and why might such evidence fail in the prediction of the clinical utility of drugs? *Annual Review of Pharmacology and Toxicology*, *55*, 169–189.
Nakai, Y., Takiguchi, T., Matsui, G., Yamaoka, N., & Takada, S. (2017). Detecting abnormal word utterances in children with autism spectrum disorders: Machine-learning based voice analysis versus speech therapists. *Perceptual and Motor Skills*, *124*, 961–973.
Niu, K., Guo, J.Y., Pan, Y.J., Gao, X., Peng, X.P., Li, N., & Li, H.L. (2020). Multichannel deep attention neural networks for the classification of autism spectrum disorder using neuroimaging and personal characteristic data. *Complexity*, *2020*. https://doi.org/10.1155/2020/1357853
Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*, 150–152.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447–453.
Ogundimu, E.O., Altman, D.G., & Collins, G.S. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*, *76*, 175–182.
Pagnozzi, A.M., Conti, E., Calderoni, S., Fripp, J., & Rose, S.E. (2018). A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective. *International Journal of Developmental Neuroscience*, *71*, 68–82.
Pan, S.J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 1345–1359.
Plitt, M., Barnes, K.A., & Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage Clinical*, *7*, 359–366.
Poldrack, R.A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry*, *77*, 534–540.
Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*, 21–45.
Preetham, P.V.S., George, F.T., George, K., & Verma, A. (2017). Deep learning based recognition of meltdown in autistic kids.
Price, W.N., 2nd & Cohen, I.G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*, 37–43.
Pulini, A.A., Kerr, W.T., Loo, S.K., & Lenartowicz, A. (2019). Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: Effects of sample size and circular analysis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*, 108–120.
Rabany, L., Brocke, S., Calhoun, V.D., Pittman, B., Corbera, S., Wexler, B.E., … & Assaf, M. (2019). Dynamic functional connectivity in schizophrenia and autism spectrum disorder: Convergence, divergence and classification. *NeuroImage Clinical*, *24*, 101966.
Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*, 1347–1358.
Ramyead, A., Studerus, E., Kometer, M., Uttinger, M., Gschwandtner, U., Fuhr, P., & Riecher-Rössler, A. (2016). Prediction of psychosis using neural oscillations and machine learning in neuroleptic-naïve at-risk patients. *The World Journal of Biological Psychiatry*, *17*, 285–295.
Raya, M.A., Giglioli, I.A.C., Marin-Morales, J., Higuera-Trujillo, J.L., Olmos, E., Minissi, M.E., … & Abad, L. (2020). Application of supervised machine learning for behavioral biomarkers of autism spectrum disorder based on electrodermal activity and virtual reality. *Frontiers in Human Neuroscience*, *14*, 90.
Razi, N.I.M., Othman, M., & Wahab, A. (2015). Classification of resting state electroencephalography for the identification of Asperger's syndrome. *Advanced Science Letters*, *21*, 3084–3087.
Riaz, A., Asad, M., Alonso, E., & Slabaugh, G. (2018). Fusion of fMRI and non-imaging data for ADHD classification. *Computerized Medical Imaging and Graphics*, *65*, 115–128.
Riaz, A., Asad, M., Alonso, E., & Slabaugh, G. (2020). DeepFMRI: End-to-end deep learning for functional

connectivity and classification of ADHD using fMRI. *Journal of Neuroscience Methods*, *335*.
Rødgaard, E.M., Jensen, K., Vergnes, J.N., Soulières, I., & Mottron, L. (2019). Temporal changes in effect sizes of studies comparing individuals with and without autism: A meta-analysis. *JAMA Psychiatry*, *76*, 1124–1132.
Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386.
Rudovic, O., Zhang, M.R., Schuller, B., & Picard, R.W. (2019). Multi-modal active learning from human data: a deep reinforcement learning approach. https://arxiv.org/abs/1906.03098
Russ, T.C., Woelbert, E., Davis, K.A.S., Hafferty, J.D., Ibrahim, Z., Inkster, B., … & Stewart, R. (2019a). How data science can advance mental health research. *Nature Human Behaviour*, *3*, 24–32.
Russ, T.C., Woelbert, E., Davis, K.A.S., Hafferty, J.D., Ibrahim, Z., Inkster, B., … & Stewart, R. (2019b). How data science can advance mental health research. *Nature Human Behaviour*, *3*, 24–32.
Rutledge, R.B., Chekroud, A.M., & Huys, Q.J. (2019). Machine learning and big data in psychiatry: toward clinical applications. *Current Opinion in Neurobiology*, *55*, 152–159.
Sadeghi, M., Khosrowabadi, R., Bakouie, F., Mandavi, H., Eslahchi, C., & Pouretemad, H. (2017). Screening of autism based on task-free fMRI using graph theoretical approach. *Psychiatry Research-Neuroimaging*, *263*, 48–56.
Sanfelici, R., Dwyer, D.B., Antonucci, L.A., & Koutsouleris, N. (2020). Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: A meta-analytic view on the state of the art. *Biological Psychiatry*, *88*, 349–360.
Schnack, H.G., & Kahn, R.S. (2016). Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Frontiers in Psychiatry*, *7*. https://doi.org/10.3389/fpsyt.2016.00050
Schooler, J.W. (2014). Metascience could rescue the 'replication crisis'. *Nature*, *515*, 9.
Schrouff, J., Rosa, M.J., Rondina, J.M., Marquand, A.F., Chu, C., Ashburner, J., … & Mourão-Miranda, J. (2013). PRoNTo: Pattern recognition for neuroimaging toolbox. *Neuroinformatics*, *11*, 319–337.
Shatte, A.B., Hutchinson, D.M., & Teague, S.J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*, 1426–1448.
Shen, M.D., Kim, S.H., McKinstry, R.C., Gu, H., Hazlett, H.C., Nordahl, C.W., … & Gu, H. (2017). Increased extra-axial cerebrospinal fluid in high-risk infants who later develop autism. *Biological Psychiatry*, *82*, 186–193.
Shen, M.D., Nordahl, C.W., Li, D.D., Lee, A., Angkustsiri, K., Emerson, R.W., … & Amaral, D.G. (2018). Extra-axial cerebrospinal fluid in high-risk and normal-risk children with autism aged 2–4 years: A case-control study. *Lancet Psychiatry*, *5*, 895–904.
Shortliffe, E.H. (1993). The adolescence of AI in medicine: Will the field come of age in the '90s? *Artificial Intelligence in Medicine*, *5*, 93–106.
Sim, I. (2019). Mobile devices and health. *New England Journal of Medicine*, *381*, 956–968.
Smith, D.M., Wang, S.B., Carter, M.L., Fox, K.R., & Hooley, J.M. (2020). Longitudinal predictors of self-injurious thoughts and behaviors in sexual and gender minority adolescents. *Journal of Abnormal Psychology*, *129*, 114–121.
Stevens, E., Atchison, A., Stevens, L., Hong, E., Granpeesheh, D., Dixon, D., & Linstead, E. (2017). A cluster analysis of challenging behaviors in autism spectrum disorder. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. (pp. 661–666). IEEE.
Stevens, E., Dixon, D.R., Novack, M.N., Granpeesheh, D., Smith, T., & Linstead, E. (2019). Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *International Journal of Medical Informatics*, *129*, 29–36.
Studerus, E., Corbisiero, S., Mazzariello, N., Ittig, S., Leanza, L., Egloff, L., … & Riecher-Rossler, A. (2018). Can neuropsychological testing facilitate differential diagnosis between at-risk mental state (ARMS) for psychosis and adult attention-deficit/hyperactivity disorder (ADHD)? *European Psychiatry*, *52*, 38–44.
Sutoko, S., Monden, Y., Tokuda, T., Ikeda, T., Nagashima, M., Funane, T., … & Dan, I. (2019). Exploring attentive task-based connectivity for screening attention deficit/hyperactivity disorder children: A functional near-infrared spectroscopy study. *Neurophotonics*, *6*, 45013.
Tariq, Q., Daniels, J., Schwartz, J.N., Washington, P., Kalantarian, H., & Wall, D.P. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLoS Med*, *15*, e1002705.
Tariq, Q., Fleming, S.L., Schwartz, J.N., Dunlap, K., Corbin, C., Washington, P., … & Wall, D.P. (2019). Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: Development and validation study. *Journal of Medical Internet Research*, *21*, e13822.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288.
Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*, 44–56.
Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*, 71–86.
Uddin, M., Wang, Y., & Woodbury-Smith, M. (2019). Artificial intelligence for precision medicine in neurodevelopmental disorders. *NPJ Digital Medicine*, *2*, 1–10.
Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J., & IEEE (2019). Kinematic features of a simple and short movement task to predict autism diagnosis. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. (pp. 1421–1424).
Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A.J. (2020). Applying machine learning to kinematic and eye movement features of a movement imitation task to predict autism diagnosis. *Scientific Reports*, *10*. https://doi.org/10.1038/s41598-020-65384-4
Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77.
Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179.
Vitiello, B. (2011). Prevention and treatment of child and adolescent depression: Challenges and opportunities. *Epidemiology and Psychiatric Sciences*, *20*, 37–43.
Voss, C., Schwartz, J., Daniels, J., Kline, A., Haber, N., Washington, P., … & Wall, D.P. (2019). Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: A randomized clinical trial. *JAMA Pediatrics*, *173*, 446–454.
Wall, D.P., Kosmicki, J., Deluca, T.F., Harstad, E., & Fusaro, V.A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry*, *2*, e100.
Walsh, C.G., Ribeiro, J.D., & Franklin, J.C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of Child Psychology and Psychiatry*, *59*, 1261–1270.
Walter, M., Alizadeh, S., Jamalabadi, H., Lueken, U., Dannlowski, U., Walter, H., … & Dwyer, D.B. (2019). Translational machine learning for psychiatric neuroimaging.

Progress in Neuro-Psychopharmacology and Biological Psychiatry, *91*, 113–121.

Wang, C., Xiao, Z., & Wu, J. (2019). Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data. *Phys Med*, *65*, 99–105.

Washington, P., Leblanc, E., Dunlap, K., Penev, Y., Kline, A., Paskov, K., … & Wall, D.P. (2020). Precision telemedicine through crowdsourced machine learning: testing variability of crowd workers for video-based autism feature recognition. *Journal of Personalized Medicine*, *10*, 86.

Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., & Kannel, W.B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*, 1837–1847.

Wingfield, B., Miller, S., Yogarajah, P., Kerr, D., Gardiner, B., Seneviratne, S., … & Coleman, S. (2020). A predictive model for paediatric autism screening. *Health Informatics Journal*, *26*, 2538–2553.

Wolfers, T., Floris, D.L., Dinga, R., van Rooij, D., Isakoglou, C., Kia, S.M., … & Beckmann, C.F. (2019). From pattern classification to stratification: Towards conceptualizing the heterogeneity of Autism Spectrum Disorder. *Neuroscience and Biobehavioral Reviews*, *104*, 240–254.

Wolfson, M., Wallace, S.E., Masca, N., Rowe, G., Sheehan, N.A., Ferretti, V., … & Burton, P.R. (2010). DataSHIELD: Resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International Journal of Epidemiology*, *39*, 1372–1382.

Wong, H.K., Tiffin, P.A., Chappell, M.J., Nichols, T.E., Welsh, P.R., Doyle, O.M., … & Tino, P. (2017). Personalized medication response prediction for attention-deficit hyperactivity disorder: learning in the model space vs. learning in the data space. *Frontiers in Physiology*, *8*, 199.

Wong, J.L., & Whitaker, D.J. (1994). The stability and prediction of depressive mood states in college students. *Journal of Clinical Psychology*, *50*, 715–722.

Woods, S.W., Choi, J., & Mamah, D. (2021). Full speed ahead on indicated prevention of psychosis. *World Psychiatry*, *20*, 223–224.

Yang, Y.J.D., Allen, T., Abdullahi, S.M., Pelphrey, K.A., Volkmar, F.R., & Chapman, S.B. (2017). Brain responses to biological motion predict treatment outcome in young adults with autism receiving Virtual Reality Social Cognition Training: Preliminary findings. *Behavior Research and Therapy*, *93*, 55–66.

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

Yassin, W., Nakatani, H., Zhu, Y.H., Kojima, M., Owada, K., Kuwabara, H., … & Koike, S. (2020). Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. *Translational Psychiatry*, *10*. https://doi.org/10.1038/s41398-020-00965-5

Yoo, J.H., Kim, J.I., Kim, B.N., & Jeong, B. (2019). Exploring characteristic features of attention-deficit/hyperactivity disorder: Findings from multi-modal MRI and candidate genetic data. *Brain Imaging and Behavior*. https://doi.org/10.1007/s11682-019-00164-x

Yoo, J.H., Sharma, V., Kim, J.W., McMakin, D.L., Hong, S.B., Zalesky, A., … Ryan, N.D. (2020). Prediction of sleep side effects following methylphenidate treatment in ADHD youth. *NeuroImage: Clinical*, *26*, 102030.

Yu, K.-H., Beam, A.L., & Kohane, I.S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*, 719–731.

Zhang, J., Yan, Y., & Lades, M. (1997). Face recognition: Eigenface, elastic matching, and neural nets. *Proceedings of the IEEE*, *85*, 1423–1435.

Zhao, Z., Zhang, X.B., Li, W.Z., Hu, X.Y., Qu, X.D., Cao, X.L., … & Lu, J.P. (2019). Applying machine learning to identify autism with restricted kinematic features. *IEEE Access*, *7*, 157614–157622.

Zhu, L., & Chang, W.K. (2019). Application of deep convolutional neural networks in attention-deficit/hyperactivity disorder classification: Data augmentation and convolutional neural network transfer learning. *Journal of Medical Imaging and Health Informatics*, *9*, 1717–1724.

Accepted for publication: 6 September 2021