Review

Biological Psychiatry: CNNI

Machine Learning for Precision Psychiatry: Opportunities and Challenges

Danilo Bzdok and Andreas Meyer-Lindenberg

ABSTRACT

The nature of mental illness remains a conundrum. Traditional disease categories are increasingly suspected to misrepresent the causes underlying mental disturbance. Yet psychiatrists and investigators now have an unprecedented opportunity to benefit from complex patterns in brain, behavior, and genes using methods from machine learning (e.g., support vector machines, modern neural-network algorithms, cross-validation procedures). Combining these analysis techniques with a wealth of data from consortia and repositories has the potential to advance a biologically grounded redefinition of major psychiatric disorders. Increasing evidence suggests that data-derived subgroups of psychiatric patients can better predict treatment outcomes than DSM/ICD diagnoses can. In a new era of evidence-based psychiatry tailored to single patients, objectively measurable endophenotypes could allow for early disease detection, individualized treatment selection, and dosage adjustment to reduce the burden of disease. This primer aims to introduce clinicians and researchers to the opportunities and challenges in bringing machine intelligence into psychiatric practice.

Keywords: Artificial intelligence, Endophenotypes, Machine learning, Null-hypothesis testing, Personalized medicine, Predictive analytics, Research Domain Criteria (RDoC), Single-subject prediction

https://doi.org/10.1016/j.bpsc.2017.11.007

Current scientific research, evaluation, and treatment in psychiatry are based on a diagnostic system solely conceived on human experiential terms, rather than on objective markers of illness. These pervasively adopted diagnostic categories have been constructed from expert opinions and enshrined in the DSM-5 and ICD-10 manuals. Yet, it is becoming increasingly clear that the pathophysiology underlying such disease definitions is rather heterogeneous (1,2). A clinically distinct mental disease is often not underpinned by an identical biology insofar as we can detect by available neuroscientific instruments. This frustration can potentially be alleviated by identifying subgroups that exhibit predictable response to treatment. However, the aspiration to automatically segregate brain disorders into natural kinds will necessitate new statistical and scientific approaches.

For decades, the dominant research paradigm to alleviate symptoms of psychiatric patients has followed an ideal chain of events: 1) initially neuroscience studies should identify new disease mechanisms (e.g., neurotransmitter pathways in animal models), 2) then innovative treatments should be explored to target the discovered disease mechanisms (e.g., design and test candidate molecular compounds), and 3) finally the new treatment should be validated by clinical trials in large cohorts (e.g., randomized clinical drug trials). Each of these three steps has frequently encountered considerable difficulties. Modern machine learning approaches may have a natural potential to improve the well-being of psychiatric patients [for excellent surveys of machine learning applications in mental health, see (3,4–6)].

One way to distinguish the diversity of quantitative analysis tools is by placing them on a continuum between classical statistics (e.g., null-hypothesis testing, Student's *t* tests, analysis of variance) and machine learning (e.g., cross-validation, support vector machines, neural network algorithms). Machine learning aims to uncover general principles underlying a series of observations without explicit instructions (7–9). Such algorithmic methods are characterized by 1) making few formal assumptions, 2) allowing the data to "speak for themselves," and 3) the ability to mine structured knowledge from extensive data. Its members include supervised methods, such as support vector machines and neural-network algorithms, specialized for best-possible outcome prediction; and unsupervised methods, such as algorithms for data clustering and dimensionality reduction, effective at discovering unknown statistical configurations in data (see Table 1 for technical terms). "Features" (traditionally called "independent variables") are fed into quantitative modeling and possibly used to predict "target variables" (traditionally called "dependent variables") (10). The recent coincidence of increasing data availability, improving computing power, and less expensive data storage has encouraged an ongoing surge in research and applications of machine learning technologies roughly since the turn of the century (11,12). As a distinctive property, new knowledge is derived by testing whether a predictive model can extrapolate patterns from one set of data to another set of data by making useful predictions in new observations (i.e., cross-validation procedures) (13). Complementing the established benefits of classical null-hypothesis testing in medicine, we will argue that

ISSN: 2451-9022

Table 1. Classes of Machine Learning Techniques With Their Statistical Purpose (in Order of Appearance)

Notion	Purpose
Supervised Learning	Models that predict a discrete outcome (e.g., healthy group vs. control group) or continuous outcome (e.g., disease severity degrees) from measures of behavior (e.g., questionnaire), brain (e.g., neural activity), or genetics (e.g., single nucleotide polymorphisms). Data have the form: features X (n subjects $\times$ p variables) and target variable y (one entry for each subject). Example: Estimate patient prognosis based on genetic profile.
Unsupervised Learning	Models that discover structure that is coherently present in the p variables across subjects. Data have the form: features X (n subjects $\times$ p variables), but no target variable y. Example: Reveal biological disease subgroups in patients based on genetic profile. Ascertaining the clinical usefulness of discovered clusters and dimensions will often require combination with supervised predictions.
Clustering	A class of unsupervised methods that uses a certain criterion to segregate a set of elements into a number of groups according to their measured similarity. Many clustering models perform hard assignments: the groups are nonoverlapping, with each element associated with only one group (i.e., "winner-takes-all" assumption). Example: k-Means clustering, hierarchical clustering, spectral clustering, density-based spatial clustering.
Dimensionality Reduction	Reexpressing observations, each quantified by many variables, in a smaller number of quintessential variables.
Support Vector Machines	A supervised model that performs prediction based on identifying observations in the data that are typical for the categories to be distinguished.
Neural-Network Algorithms	A supervised model that performs prediction based on a nonlinear, multilayer variant of linear regression. "Deep" neural networks are a modern version with a higher number of nonlinear processing layers.
Cross-validation	A two-step procedure used as the de facto standard to estimate the capacity of a pattern learning model to extrapolate to future data samples. First, the predictive model is fitted on training data and, second, its generalization performance is evaluated on test data (out of sample). The process is repeated for different splits of the data (often 5 or 10 times).
In-Sample Estimate	Prediction performance measured in the same data that was also used to fit the model. Example: Most applications of linear regression in biomedical research exclusively compute in-sample estimates, without considering out-of-sample estimates.
Out-of-Sample Estimate	Prediction performance measured in new data that was not used to fit a model. Example: In machine learning, it is the core metric of how successful extrapolation of a derived pattern to new, independent data is quantified.
Training Data	A model is fitted to identify a certain pattern from a larger part of the available data.
Test Data	An already fitted model is used for prediction in a smaller part of the available data.
Multiclass Learning	Applying a supervised model to predict an outcome y that denotes more than two (possibly hundreds of) categories. Example: Model predicts best among (many) more than two drug treatment options.
Multitask Learning	Applying a supervised model to simultaneously predict several outcomes $y_1, y_2, \ldots, y_m$. Example: Model uses the same brain scans to conjointly predict drug treatment options, candidate diagnoses, and disease trajectories.
Manifolds	Effective reexpression of data by revealing distinct factors that collectively underlie a set of observations. Example: Everyday objects are manifolds in a 3-dimensional space (e.g., a flower), although there is a variety of perspectives from which humans can gather and contemplate information about an object, including vision, audition, touch, smell, taste, and many others.
Representation Learning	Applying models that can automatically extract hidden manifolds from data.
Latent Factor Modeling	A class of unsupervised methods that use a certain criterion to stratify a set of elements with their respective relationships to a number of hidden components of variation so as to maximize between-component dissimilarity. Many latent factor models perform soft assignments: component of variations are overlapping, with each element associated to each component to a certain extent (i.e., no "winner-takes-all" assumption). Example: Latent Dirichlet allocation, autoencoders, nonnegative matrix factorization, isomap, t-distributed stochastic neighbor embedding
k-Means Clustering	A popular clustering model that partitions the p variables of X into k nonoverlapping groups. Example: Use genetic information to group mammals into human and nonhuman primates ($k = 2$).
Hierarchical Clustering	A popular clustering model that builds a nested tree by successively partitioning the p variables of X into k always more fine-grained nonoverlapping groups. All clustering solutions from $k = 1$ group to $k = n$ groups are often computed. Example: Use genetic information to group mammals ($k = 1$) into human and nonhuman primates ($k = 2$), which are then grouped into humans, apes, and monkeys ($k = 3$), and so forth.
Latent Dirichlet Allocation	A latent factor model that stratifies countlike data into overlapping components of variation. Example: Extract coherent combinations of number of times (positive discrete numbers) words occurred during an unstructured clinical interview.
Autoencoders	A latent factor model that stratifies continuous data into overlapping components of variation. Example: Extract coherent combinations of item scores from a structured clinical questionnaire (positive or negative nondiscrete numbers).

Table 1. Continued

Notion	Purpose
Dictionary Learning	Superclass of many clustering models and latent factor models that try to extract a set of atomic representations (i.e., the dictionary) from a set of observations that variably add up to each specific observation.
Overfitting	The model fits the data overly well, at the expense of generalization performance. Intuitively, the model "hallucinates" relevant patterns in the data. The more complex the patterns that can be learned by a model, the bigger the danger of overfitting.

Application examples for using machine learning in mental health are available elsewhere (3–6).

machine learning is predisposed to address many challenges in the upcoming era of precision psychiatry.

OPPORTUNITIES

Current drug treatment choices are only successful in roughly half of patients (14), and similar considerations apply to psychotherapy (15). In fact, the psychiatrist's choice of the best-possible treatment option often does not depend on knowledge of what has caused or maintains the mental disease of a given patient. A research and treatment strategy that does not depend on full understanding of complex disease mechanisms may be less expensive and incur shorter delays between bench and bedside (16). Systematically benchmarking the predictability of clinical quantities in single patients could improve clinical symptoms faster and reduce subjective suffering in many mental diseases. Even moderately successful predictive models can be highly useful in clinical practice (3). This is because of the unfortunate normal case of trial-and-error treatment with psychotropic drugs and other types of treatment for many mental diseases (17). While the traditional research goal was to introduce novel treatment options that benefit some majority of a particular clinical group, an attractive alternative research goal is to improve the choice of existing treatment options by predicting their effectiveness in single patients. More and more studies now indicate that a specific drug or psychotherapy treatment can be successful in a certain patient subgroup and unsuccessful in another patient subgroup labeled with the identical diagnosis [for an overview, see (18)]. In a successful example of using machine learning in psychiatry, the discovered patient subgroups could indeed be used to predict which patient would profit from brain-stimulation treatment (19). This questions the primacy of drawing conclusions on the group-level and opens the possibility of building objective algorithmic frameworks with individual treatment-response prediction across a diversity of psychiatric conditions.

Machine learning offers a set of tools that are particularly suited to achieve individual-level clinical predictions. Predictive models are conceptually positioned between genetic risk variants as an individual's blueprint at the one extreme and clinical symptoms as an individual's behavioral manifestations at the other extreme. Benefitting from a variety of intermediate phenotypes has the translational potential to refine clinical management by early diagnosis and disease stratification, selection between drug treatments, treatment adjustment, and prognosis for psychiatric care tailored to each individual (19). Learning algorithms can be directly applied in single patients to predict inherently valid and immediately useful clinical objects (5), such as choosing drug dosage. There are a number of reasons why many machine learning methods are naturally applicable for prospective clinical predictions on the single-subject level, whereas the currently most widespread statistical methods may be more tuned to group-level analysis.

Focus on Prediction

Machine learning methods have a long-standing focus on prediction as a metric of statistical quality (10). Support vector machines, neural-network algorithms, and many other trained predictive models are readily able to estimate an outcome from only one observation, such as when querying answers from behavioral, neural, or genetic measurements of a single patient (3,4). In contrast, classical statistical methods are often used in medical research to explain variance of and formally test for group effects. Analysis of variance, Student's *t* test, and many other commonly used tools grounded in the notion of statistical significance have a less obvious ability for judgments on one specific individual in a group. Thus, common routines of machine learning and classical statistics serve rather distinct statistical purposes. The two statistical cultures perform different types of principled assessment for successful extrapolation of an effect beyond the data at hand that are rooted in different mathematical contexts. As an important practical consequence, machine learning and classical statistics do not judge data on the same aspects of evidence: an observed effect assessed to be statistically significant by a *p* value does not in all cases yield a high prediction accuracy in new, independent data, and vice versa (4,8,20,21).

Empirical Model Evaluation

By quantifying the prediction success in new individuals (so-called out-of-sample estimates) many machine learning approaches naturally adopt a prospective viewpoint and can directly yield a notion of clinical relevance. Instead, classical approaches based on null-hypothesis testing often take a retrospective flavor as they usually revolve around finding statistical effects in the dataset at hand (so-called in-sample estimates) based on prespecified modeling assumptions, typically without explicitly evaluating some fitted models on unseen or future data points (20). Hence, ubiquitous techniques for out-of-sample generalization in machine learning are likely candidates for enabling a future of personalized psychiatry. This is because predictive models can be applied to and obtain answers from a single patient.

Two-Step Procedures

Traditional null-hypothesis testing takes the form of a one-step procedure. That is, the whole dataset is routinely used to produce a *p* value or an effect-size measure in a single

process. An obtained *p* value or effect size can itself not be used to judge other data in some later step. In contrast, machine learning models are typically evaluated by cross-validation procedures as a gold standard to quantify the ability of a learning algorithm to extrapolate beyond the dataset at hand (10). In a two-step procedure, a learning algorithm is fitted on a bigger amount of available data (so-called training data) and the ensuing "trained" learning model is empirically evaluated by application to a smaller amount of new data (so-called test data). This two-step nature of machine learning workflows lends itself particularly well to, in a first step, extract structured knowledge in large openly available or hospital-provided datasets. In a second step, the ensuing trained predictive models can be shared collaboratively as a research product (6) and be applied with little effort in a possibly large number of individual patients in various mental health contexts.

Suited to Observational Data

Many methods from classical statistics have probably been devised for experimental data that are acquired in a context where a set of target variables has been systematically manipulated by the investigator (e.g., randomized clinical trials with placebo group and active treatment group). However, precision medicine in psychiatry is likely to exploit especially observational data (e.g., blood and metabolic samples, movement and sleeping patterns, electroencephalograms, brain scans, and genetic variants) that were acquired without a carefully controlled influence in an experimental setup and to which machine learning tools may be more closely tuned [e.g., (19,22,23)].

Handle Many Outcomes at Once

Machine learning is also a pertinent choice for comparisons between many (potentially hundreds of) possible diagnoses and other multioutcome settings. Classical significance testing is probably most often used to decide between two possible outcomes, expressed in the null and alternative hypothesis, by considering the probability of obtaining an equal or more extreme effect in the data under the null hypothesis (24). This is often used in group analysis to formally determine a scientifically relevant difference between healthy subjects (i.e., typically corresponding to the null hypothesis) and psychiatric patients as defined by a DSM or ICD category (i.e., typically the alternative hypothesis) or when comparing a placebo treatment (i.e., null hypothesis) against a new treatment (i.e., alternative hypothesis). In everyday practice in psychiatry, the more challenging question is typically not whether a patient has a mental disease, but the differential diagnosis between a number of likely disease categories—the transdiagnostic setting. Analogously, whether a patient needs treatment is routinely an easier clinical decision than choosing between numerous competing treatment options. Treatment response prediction is rarely a binary yes-or-no decision and requires consideration of several treatment options in the same statistical analysis.

Machine learning is well suited to this goal in the form of multiclass prediction and multitask learning (23,25,26), unlike many approaches for statistical significance assessment and tests of group differences. Most machine learning approaches that are applicable when aiming to distinguish two groups or two treatment options can be extended to consider a wide range of possible outcomes. For instance, quantitative brain measurements from one patient can be fed into prediction models to simultaneously infer a probabilistic stratification over several differential diagnoses, many candidate treatment options, risk outcomes, and possible long-term clinical prognoses (e.g., full recovery vs. partial residuals vs. severe chronic illness). Additionally, applying learning algorithms to compare patients versus control subjects does not allow for evaluating how specific an achieved prediction is for the given psychiatric group (6). Besides the advantage of replacing artificial, mutually exclusive dichotomies by predicting several outputs in concert, the prediction accuracy often improves when statistical strength can be shared between the variation in the data associated with the respective outcomes (27). That is, statistical estimation of different predictions can paradoxically become easier when considered in parallel by one model instead of answering isolated statistical questions. In sum, there are clear incentives and readily applicable statistical tools to go beyond group-level comparisons à la normal versus diseased (28,29). Importantly, machine learning is naturally suited for choosing between a potentially massive number of possible options in a single patient and ranking the possibilities according to pertinence. This is the case in the transdiagnostic setting where the pertinence of several psychiatric diagnoses needs to be predicted hand in hand by one statistical model.

Explore Manifolds in Complex Data

Besides the intricacies of considering several diagnostic categories at once, the diagnostic categories themselves have repeatedly been called into question due to their lack of neurobiological validity and clinical predictability (1,2). The disease definitions cataloged in the DSM and ICD manuals do not always align well with new behavioral, neuroscientific, and genetic evidence (Figure 1). Psychiatric disorders have been defined in the DSM and ICD with a focus on ensuring effective communication of diagnoses between clinicians (i.e., interrater reliability) rather than the goal to capture natural kinds in biological reality (1). Autism, schizophrenia, and an increasing number of other psychiatric diseases are suspected to be spectrum disorders—heterogeneous etiological and pathophysiological factors being summarized under the same umbrella term (30,31). This conceptualization is also more compatible with a smooth transition between healthy and psychiatrically diagnosed individuals. Machine learning offers a rich variety of tools that lend themselves to endophenotype modeling.

Hence, among many clinicians and researchers, there is a growing wish to supplement discrete disease definitions in the form of categories with a continuous, dimensional symptom system. To satisfy the need to cut across diagnostic boundaries, the Research Domain Criteria (RDoC) initiative (32) has been launched as a translational program to elucidate the hidden structure underlying psychopathology. By synergistic integration of self-reports, neuropsychological tests, brain measurements, and genetic profiles, RDoC wants to "better understand basic dimensions of functioning . . . from normal to abnormal" (National Institute of Mental Health, Bethesda, MD;

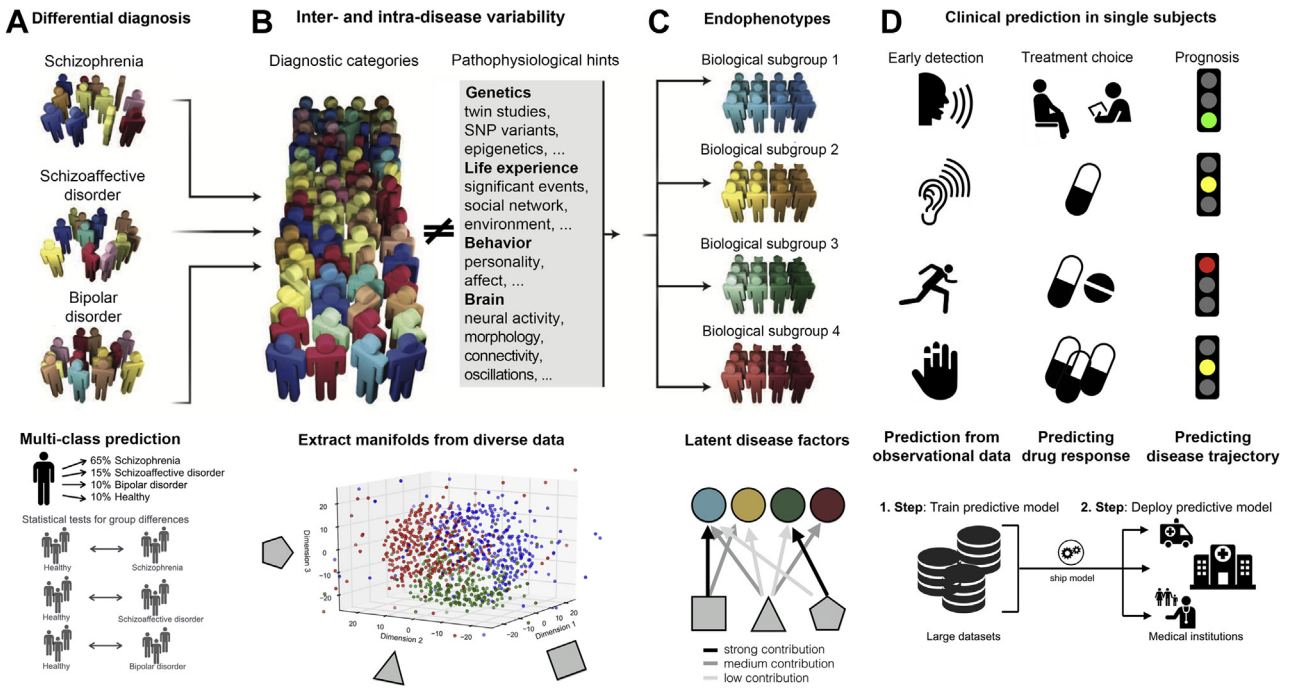

Figure 1. Current challenges for precision medicine in psychiatry and possible solutions from machine learning. **(A)** Basic and clinical psychiatric research frequently investigates a given patient population by group comparison against the healthy population, possibly creating artificial dichotomies. Many machine learning approaches can be extended to compare observations from a number of groups in the same statistical estimation (i.e., multiclass prediction; see also Table 1). **(B)** The diagnostic categories in the ICD and DSM manuals were primarily designed to reliably describe symptom phenomenology and are frequently incongruent with new behavioral, neural, and genetic research evidence. Machine learning methods can automatically extract currently unknown patterns of variation in individuals simultaneously from heterogeneous data that cut across traditional diagnoses (i.e., manifolds). **(C)** Assigning a patient to only one diagnostic category may ignore that different pathophysiological mechanisms (i.e., endophenotypes) can contribute to the same clinical picture. Instead of relying on categorical assignments, biologically defined subgroups can be described by continuous contributions of several disease processes in graded degrees (i.e., latent factor models). **(D)** Psychiatric care today often resorts to trial and error. Predictive models could improve patient care by earlier detection, treatment selection and adjustment, and inference of disease trajectory. After a machine learning algorithm has been trained in extensive data (i.e., in sample), the trained predictive model can be used for personalized prediction without database access (i.e., out of sample). SNP, single nucleotide polymorphism. [Reprinted with permission and modified from Insel and Cuthbert (2).]

https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml) without relying on presupposed disease definitions. The discovered fundamental dimensions of behavior and its disturbances are expected to motivate new research approaches aimed at reformating psychiatric nosology. RDoC thus recommends going from scientific evidence to organically deriving new disease factors. This framework thus contrasts the dominant agenda in psychiatric research that goes from disease categories defined based on the DSM and ICD to generating scientific evidence. The RDoC approach is conceptually compatible with the fact that psychiatric patients exhibit clusters of psychopathological symptoms and that many symptoms are shared among, rather than being unique to, different psychiatric disorders. RDoC is also naturally compatible with the accumulating evidence that risk alleles are partly shared between psychiatric disorders (33), while different sets of risk alleles can lead to an identical psychiatric phenotype (34).

Of note, this renewed focus on fundamental building blocks of mental disturbance finds a direct correspondence in the emphasis of the machine learning community on representation learning for discovering hidden structure in complex data (35). In particular, the multidimensional conception underlying the RDoC initiative is reminiscent of the notion of manifolds that is common in the machine learning field (7). In a setting with possibly many high-resolution measurements (brain scans, sequenced genome, etc.), a manifold describes coherent low-dimensional directions of relevant variation in the data. Here, members of a coherent class would be expressed as "a set of points associated with a neighborhood around each point" (7). In psychiatry, the manifold notion corresponds to the hope that the nature of psychiatric disorders and their complex relationships could be described effectively in a small number of hidden dimensions: each a distinct direction of variation in heterogeneous data sources. Variation captured across behavioral, experiential, neural, and genetic measurements with tens of thousands of input variables can hopefully be effectively expressed along a manifold that concerns only a much smaller number of yet-to-be-discovered disease dimensions.

Indeed, supplementing traditional null-hypothesis testing, machine learning has a rich legacy of algorithm developments that can now be repurposed to automatically extract from data manifolds that describe behavior, life experience, brain, or genetics. Representation-learning algorithms operate on the assumption that the measured data have been generated by a

set of underlying constituent factors. Unfortunately, however, many traditional clustering algorithms, such as hierarchical and *k*-means clustering, assign each individual exclusively to only one group ("winner-takes-all" assumption). These biologically and clinically implausible statistical assumptions can be relaxed by recourse to latent factor models (7), including latent Dirichlet allocation (22), autoencoders (36), and many other dictionary-learning procedures. Latent factor models can uncover an underlying manifold of hidden directions of variation by assigning each individual to each of the groups to different degrees. Technically, the same manifold dimension, reflecting a distinct disease process, is thus allowed to contribute in sophisticated and nuanced ways to several psychiatric disorders with clinical pictures as diverse as schizophrenia, autism, and bipolar disorder, in line with the RDoC agenda. Thus, given the prevailing lack of objective markers in psychiatry, there is merit in revealing, formalizing, and clinically exploiting currently unknown interindividual variation.

CHALLENGES

There are many good reasons for extending machine learning applications to psychiatric research and practice. The various pitfalls in their everyday usage have been previously outlined in machine learning in general [e.g., (37,38)] and in neuroscience in particular [e.g., (5,39–41)]. Here, we will focus on more general obstacles that we need to overcome.

Reproducibility

Prototyping, iteratively improving, and benchmarking machine learning pipelines involves many complicated, interdependent choices. Such multistep workflows are becoming challenging to fine-tune manually. The increasing flexibility of analysis pipelines is raising the concern that obtained findings might less reliably replicate in later studies (42). Successful deployment of predictive models on the clinical ward may profit from seamless exchange of predictive models in the research community. The final predictive models should pass the prospective test of handling new subjects in other research laboratories and should persist across data-acquisition means (e.g., Siemens/Philips/GE brain scanners), across geographic locations (e.g., United States, Europe, and Asia), and across populations (e.g., same mental disorder with different comorbidity profiles), as well as persist for different success metrics (e.g., sensitivity and specificity) and clinical settings (e.g., rural practitioner vs. university hospital) (6). Moreover, clinical biomarkers derived from genetics or neuroimaging will potentially be accredited through randomized clinical trials.

Data Availability

The primary limitation for deploying state-of-the-art algorithms to personalize psychiatric care is probably the size of today's datasets (i.e., number of subjects) and their insufficient phenotypic detail (e.g., medical history, comorbidities, progression in symptoms, treatment, and response). In fact, "mental health captures arguably the largest amount of data of any medical specialty" (43). However, compared with some nonmedical domains, psychiatric research is still far from the >1,000,000 examples threshold where the predictive power of highly successful, data-demanding models has been showcased (7). Small sample sizes exacerbate the tendency of adaptive models learning noise in the data—overfitting (44). Besides limited data quantity, exploiting emerging machine learning technologies is hindered by the insufficient specificity and granularity of the participants' behavioral information. First, many phenotypes of interest do not vary enough across subjects in general-purpose datasets. Second, deploying emerging predictive algorithms for successful subject-level prediction of practically useful clinical endpoints will probably depend on datasets with rich and meticulously acquired patient documentation.

Data Management

Over the last 10 years, growing sample sizes were enabled by national and international consortia that accumulate, curate, and distribute data across research groups, including Autism Brain Imaging Data Exchange and Alzheimer's Disease Neuroimaging Initiative. An important prerequisite is the willingness to embrace the values and habits of the open-science movement (45). While some investigators deem restricted access to research data unethical, data sharing also invokes privacy concerns (46), such as recognizing a participant's face in anatomical brain scans. In general, agreement on machine-readable data structures will become increasingly useful for machine learning applications (47).

Heterogeneous and Incomplete Data

A first generation of data initiatives (e.g., Autism Brain Imaging Data Exchange, Alzheimer's Disease Neuroimaging Initiative, ENIGMA) were retrospective collections of independently acquired data from different clinical centers. Such data repositories frequently vary in data quality, acquisition parameters, hardware and software versions, preprocessing, artifacts, used psychological assessments, and missing data. Across-site heterogeneity might explain why, counterintuitively, predictive model performances have been repeatedly reported to decrease as the available data increase (6). A second generation of data initiatives (e.g., Human Connectome Project and UK Biobank) first decided on common data acquisition practices and then coordinated distributed data collection. Ensuing repositories offer higher data comparability owing to efforts including calibrated acquisition conditions, staff training, or traveling experts. Of note, homogenizing data acquisition and analysis can maximize group differences and alleviate confound problems, whereas homogenizing population samples may not be optimal in all cases. A balance must be found between conservative manual selection of samples with convincing model performance and liberal samples more representative of clinical reality.

Longitudinal Data

Many mental disorders have a characteristic time-varying nature. Retrospective data collections typically lend themselves more to cross-sectional analysis, while prospectively collected data can be more suitable for longitudinal analysis. Most machine learning approaches and their clinical applications currently focus on cross-sectional findings. Computational psychiatric research may bear a blind spot regarding disease trajectories and longer-term health outcomes (18). A

promising avenue to accumulate massive longitudinal data may be offered by technical devices carried by subjects (48). For instance, voice data from smartphones could enable early detection of health care events, such as thought disorders, depressive episodes, or suicide attempts. More generally, digital sensors are entering everyday life [see "Internet of Things" in (48)] and can continuously monitor diverse behaviors, including sleep patterns, communication habits, gait, and geographical movement. This may enable continuously improving machine learning models.

Confounding

Accumulating observational human data is often less expensive and easier, while the lack of experimental protocols exacerbates control of confounding influences (49,50). Essentially, the prediction performance becomes inflated if the training data used for model fitting and the testing data are somehow statistically dependent, even if they are contaminated in subtle ways. Researchers are challenged to identify and account for influences unintentionally contributing to high prediction accuracies, including age, gender, culture, smoking, caffeine, drug use, and physiological noise (e.g., respiration and heart beat). Sociologically, bias may inadvertently arise because clinical research typically recruits subjects with exposure to psychiatric institutions, rather than never-diagnosed individuals with mental problems. For instance, high-functioning, subclinical individuals with schizophrenia, never in contact with a psychiatrist, might systematically evade research efforts.

CONCLUSIONS

The soaring costs of psychiatric disease prompt a global challenge for our societies (51). Whether personalized medicine can be realized to enhance psychiatric care is largely a statistical question at its heart. For many decades, "the group" has served mental health investigators as the primary working unit. Facilitated acquisition of always more detailed and diverse information on psychiatric patients is now bringing another working unit within reach—the single patient. Rather than preassuming existence of disease categories and formally verifying prespecified neurobiological hypotheses, an increasingly attractive alternative goal is to let the data guide the investigation. Following the growing data richness and changing research questions, some long-trusted statistical methods may be superseded as the best tool in the box. The statistical properties of learning algorithms could thus enable clinical translation of empirically justified single-patient prediction in a fast, cost-effective, and pragmatic manner.

For a long time, knowledge generation in basic neuroscience and clinical decision-making in psychiatry have been grounded in classical statistics with formal tests for group differences in frequently small samples. However, machine learning methods may be particularly tuned to the ambitions of precision psychiatry because they can directly translate complex pattern discovery in "big data" into clinical relevance. For most learning algorithms, it is standard practice to estimate the generalization performance to other samples by empirically cross-validating the trained algorithms on fresh data: in this case, individual subjects. This stands in stark contrast to classical statistical inference that seeks to reject the null hypothesis by considering the entirety of a data sample (24): in this case, all available subjects. In the hypothesis-testing framework, the desired relevance of a statistical relationship in the general population is ensured by formal mathematical proofs and is not commonly ascertained by explicit model evaluation on independent data (8,24).

From a larger perspective, it is particularly challenging to verbalize mechanistic hypotheses for psychiatric disorders at the most pertinent abstraction level, ranging from molecular histone-tail methylation in the cell nucleus to urbanization trends in society as a whole. This epistemological challenge highlights more human-independent pattern learning algorithms as an underexploited research avenue. Learning algorithms can automatically identify disease-specific biological aspects that achieve intrinsically valid and immediately useful clinical predictions. Ultimately, by allying with recent statistical technologies, we may more effectively impact mental disease that arises at the interplay between genetic endowment and life experience—both of which are unique to each individual.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by the German Research Foundation (DFG, BZ2/2-1, BZ2/3-1, and BZ2/4-1 to DB), International Research Training Group (Grant No. IRTG2150 to DB), Amazon AWS Research Grant (2016 and 2017) to DB, the German National Academic Foundation (DB), and the START-Program of the Faculty of Medicine, Rheinisch-Westfälische Technische Hochschule Aachen University (DB).

We thank Bertrand Thirion (French Institute for Research in Computer Science and Automation [INRIA]), Teresa Karrer (RWTH), Julius Kernbach (RWTH), Dominic Dwyer (Ludwig Maximilian University of Munich), Cedric Xia (University of Pennsylvania), and Efstathios Gennatas (University of Pennsylvania) for valuable comments on a previous version of the manuscript.

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Psychiatry (DB), Psychotherapy and Psychosomatics, RWTH Aachen University; JARA-BRAIN (DB), Jülich-Aachen Research Alliance, Aachen; Department of Psychiatry and Psychotherapy (AM-L), and Bernstein Center for Computational Neuroscience Heidelberg-Mannheim (AM-L), Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany; and Parietal team (DB), INRIA, Neurospin, Gif-sur-Yvette, France.

Address correspondence to Danilo Bzdok, M.D., Ph.D., Research Center Jülich, Institute of Neuroscience and Medicine 1, Pauwelsstraße 30, Aachen 52074, Germany; E-mail: danilo.bzdok@rwth-aachen.de.

Received Oct 17, 2017; accepted Nov 17, 2017.

REFERENCES

1. Hyman SE (2007): Can neuroscience be integrated into the DSM-V? Nat Rev Neurosci 8:725–732.
2. Insel TR, Cuthbert BN (2015): Brain disorders? Precisely. Science 348:499–500.
3. Stephan KE, Schlagenhauf F, Huys QJM, Raman S, Aponte EA, Brodersen KH, *et al.* (2017): Computational neuroimaging strategies for single patient predictions. Neuroimage 145:189–199.
4. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017): Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage 145:137–165.
5. Huys QJM, Maia TV, Frank MJ (2016): Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci 19:404–413.

6. Woo C-W, Chang LJ, Lindquist MA, Wager TD (2017): Building better biomarkers: Brain models in translational neuroimaging. Nat Neurosci 20:365–377.
7. Goodfellow IJ, Bengio Y, Courville A (2016): Deep Learning. Cambridge, MA: MIT Press.
8. Breiman L (2001): Statistical modeling: The two cultures. Stat Sci 16:199–231.
9. Bzdok D (2017): Classical statistics and statistical learning in imaging neuroscience. Front Neurosci 11:543.
10. Hastie T, Tibshirani R, Friedman J (2001): The Elements of Statistical Learning. Springer Series in Statistics. Heidelberg, Germany: Springer.
11. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, *et al.* (2011): Big data: The next frontier for innovation, competition, and productivity. Technical report. Washington, DC: McKinsey Global Institute.
12. Efron B, Hastie T (2016): Computer-Age Statistical Inference. Cambridge, UK: Cambridge University Press.
13. Shalev-Shwartz S, Ben-David S (2014): Understanding Machine Learning: From Theory to Algorithms. Cambridge, UK: Cambridge University Press.
14. Wong EHF, Yocca F, Smith MA, Lee C-M (2010): Challenges and opportunities for drug discovery in psychiatric disorders: The drug hunters' perspective. Int J Neuropsychopharmacol 13:1269–1284.
15. Hofmann SG, Asnaani A, Vonk IJ, Sawyer AT, Fang A (2012): The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognit Ther Res 36:427–440.
16. Perna G, Nemeroff CB (2017): Personalized medicine in psychiatry: Back to the future. Pers Med Psychiatry 1:1.
17. Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, *et al.* (2006): Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR* D report. Am J Psychiatry 163:1905–1917.
18. Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S (2015): Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85:11–26.
19. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, *et al.* (2017): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23:28–38.
20. Bzdok D, Yeo BTT (2017): Inference in the age of big data: Future perspectives on neuroscience. Neuroimage 155:549–564.
21. Shmueli G (2010): To explain or to predict? Stat Sci 25:289–310.
22. Zhang X, Mormino EC, Sun N, Sperling RA, Sabuncu MR, Yeo BTT (2016): Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. Proc Natl Acad Sci U S A 113:E6535–E6544.
23. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al.* (2017): Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118.
24. Wasserstein RL, Lazar NA (2016): The ASA's statement on p-values: Context, process, and purpose. Am Stat 70:129–133.
25. Breiman L, Friedman JH (1997): Predicting multivariate responses in multiple linear regression. J R Stat Soc Ser B Stat Methodol 59:3–54.
26. Caruana R (1998): Multitask Learning: Learning to Learn. New York: Springer, 95–133.
27. Caruana RA (1993): Multitask connectionist learning. In: Proceedings of the 1993 Connectionist Models Summer School. Hillsdale, NJ: Lawrence Erlbaum, 372–379.
28. Ransohoff DF, Feinstein AR (1978): Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299:926–930.
29. Perlis RH (2011): Translating biomarkers to clinical practice. Mol Psychiatr 16:1076–1087.
30. Cuthbert BN, Insel TR (2013): Toward the future of psychiatric diagnosis: The seven pillars of RDoC. BMC Med 11:126.
31. Van Os J (2016): "Schizophrenia" does not exist. BMJ 352:i375.
32. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, *et al.* (2010): Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. Am J Psychiatry 167:748–751.
33. Cross-Disorder Group of the Psychiatric Genomics C (2013): Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. Lancet 381:1371–1379.
34. Ehrenreich H, Mitjans M, Van der Auwera S, Centeno TP, Begemann M, Grabe HJ, *et al.* (2018): OTTO: a new strategy to extract mental disease-relevant combinations of GWAS hits from individuals. Mol Psychiatry 23:476–486.
35. Bengio Y, Courville A, Vincent P (2013): Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828.
36. Hinton GE, Salakhutdinov RR (2006): Reducing the dimensionality of data with neural networks. Science 313:504–507.
37. James G, Witten D, Hastie T, Tibshirani R (2013): An Introduction to Statistical Learning. New York: Springer.
38. Domingos P (2012): A few useful things to know about machine learning. Commun ACM 55:78–87.
39. Pereira F, Mitchell T, Botvinick M (2009): Machine learning classifiers and fMRI: A tutorial overview. Neuroimage 45:199–209.
40. Haynes J-D (2015): A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. Neuron 87:257–270.
41. Mur M, Bandettini PA, Kriegeskorte N (2009): Revealing representational content with pattern-information fMRI—an introductory guide. Soc Cogn Affect Neurosci 4:101–109.
42. Squeglia LM, Ball TM, Jacobus J, Brumback T, McKenna BS, Nguyen-Louie TT, *et al.* (2017): Neural predictors of initiating alcohol use during adolescence. Am J Psychiatry 174:172–185.
43. Eyre HA, Singh AB, Reynolds C (2016): Tech giants enter mental health. World Psychiatry 15:21–22.
44. Varoquaux G (2017): Cross-validation failure: Small sample sizes lead to large error bars [published online ahead of print Jun 24]. Neuroimage.
45. Longo DL, Drazen JM (2016): Data sharing. N Engl J Med 374:276–277.
46. Poldrack RA, Gorgolewski KJ (2014): Making big data open: Data sharing in neuroimaging. Nat Neurosci 17:1510–1517.
47. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, *et al.* (2016): The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data 3, 160044.
48. Manyika J, Chui M, Bisson P, Woetzel J, Dobbs R, Bughin J, *et al.* (2015): Unlocking the Potential of the Internet of Things. Washington, DC: McKinsey Global Institute.
49. Weinberger DR, Radulescu E (2015): Finding the elusive psychiatric "lesion" with 21st-century neuroanatomy: A note of caution. Am J Psychiatry 173:27–33.
50. O'Neil C (2016): Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown.
51. Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, *et al.* (2011): Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 21:718–779.