Machine learning and big data in psychiatry: toward clinical applications
Robb B Rutledge¹,², Adam M Chekroud³,⁴ and Quentin JM Huys¹,⁵,⁶

Psychiatry is a medical field concerned with the treatment of mental illness. Psychiatric disorders broadly relate to higher functions of the brain, and as such are richly intertwined with social, cultural, and experiential factors. This makes them exquisitely complex phenomena that depend on and interact with a large number of variables. Computational psychiatry provides two ways of approaching this complexity. Theory-driven computational approaches employ mechanistic models to make explicit hypotheses at multiple levels of analysis. Data-driven machine-learning approaches can make predictions from high-dimensional data and are generally agnostic as to the underlying mechanisms. Here, we review recent advances in the use of big data and machine-learning approaches toward the aim of alleviating the suffering that arises from psychiatric disorders.

Addresses
¹ Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, England, United Kingdom
² Wellcome Centre for Human Neuroimaging, University College London, London, England, United Kingdom
³ Department of Psychiatry, Yale University, New Haven, CT, United States
⁴ Spring Health, New York, NY, United States
⁵ Division of Psychiatry, University College London, London, England, United Kingdom
⁶ Camden and Islington NHS Foundation Trust, London, England, United Kingdom

Corresponding author: Huys, Quentin JM (qhuys@cantab.net)

Individual predictions through machine-learning approaches

Other areas of medicine have been successful in developing approaches that more efficiently use existing data. The obvious question is what type of data is useful. Psychiatry’s first port of call should be existing data from randomized controlled trials (RCTs), since patients are usually phenotyped carefully and monitored closely over time. However, RCTs are expensive to run, and thus usually only enroll a relatively small number of patients. Statistically speaking, datasets with more questionnaire items than participants can pose problems regarding overfitting; when there are more predictors than data points, enough predictors can always (please emphasize) produce perfect predictions [6]. Even standard cross-validation approaches can overestimate predictive accuracy and reduce the true generalizability of models. Nevertheless, Machine learning provides ways to address this (Box 1), with promising results.

Uher and co-workers first extracted robust factors from questionnaire responses in an unsupervised manner and then tested their power to predict treatment outcomes [7,8]. Other subsequent machine-learning studies including neuroimaging studies (e.g. [9]) focused on recovering diagnostic information from high-dimensional data. Recently, large numbers (>100) of clinical variables have been successfully used to predict treatment response. In first-episode psychosis, response to five different...
Antipsychotic treatments can be predicted from questionnaire responses with 71% balanced accuracy across treatment sites [10]. Antidepressant response can be predicted using a similar approach [11]. When other forms of data have been included in predictive models, performance has generally improved modestly beyond clinical and sociodemographic data. Whelan et al. predicted development of alcohol misuse in adolescents (n = 692) from brain structure and function, individual personality and cognitive differences, environmental factors including gestational cigarette and alcohol exposure, life experiences, and candidate genes [12]. While items relating to patient history and personality had the greatest predictive value, neuroimaging and genetic variables still contributed some unique information. Other studies — although small — also find that cognitive markers and structural and functional MRI can predict clinical outcomes [13–17]. EEG-derived resting-state rostral cingulate theta activity predicts antidepressant response beyond other variables ([18] but see Ref. [19]).

As well as predicting how likely an individual is to recover overall, it is also useful to understand whether certain treatments are more likely to work than others. Some progress has been made, for example, in contrasting antidepressants versus psychotherapy (e.g. [20]). One other approach developed De Rubeis et al. [21] based on work by Barber and Muenz [22] developed predictors for each RCT arm separately and then applied them to both arms. The difference in predicted outcome can then be interpreted as a reflection of how likely the individual is to respond to one treatment versus the other [23]. This was also attempted with a broad array of clinical, neuroimaging, and behavioral variables (c.f [24]). Critics have argued that these tools are not yet useful for clinical practice [25–29] because their performance remains modest in terms of explained variance and predictive accuracy [15,30], and the studies have enrolled very small samples relative to the amount of data used for prediction [31]. Nevertheless, one promising study (Figure 1a) suggests that resting-state EEG after medication washout can be used to make antidepressant treatment recommendations to clinicians [32,33] and merits efforts for replication.

Finally, applying machine learning to existing data might also help identify illness phenotypes that relate more closely to the efficacy of existing treatments. Two recent studies offer notable examples. Applying hierarchical clustering to over 7000 depressed patients from multiple RCTs, Chekroud et al. described how symptoms in common depression rating scales hang together, finding three statistically robust and replicable symptom clusters: a mood/emotional cluster, a sleep/insomnia cluster, and an atypical symptom cluster [34]. The three clusters differed in their responsiveness to antidepressants, suggesting that treatments could be targeted in individuals based on symptoms. Drysdale et al. employed similar hierarchical clustering to resting-state fMRI data from over 300 depressed patients [35], identifying four clusters of patients that differed in their responsiveness to transcranial magnetic stimulation therapy. Both studies share commonalities that proved successful: an approach for parsing phenotypic heterogeneity, aggregation of data sets large enough to support the statistical techniques, and demonstration that identified clusters meaningfully relate to treatment outcomes. These large-scale symptom-clustering and patient-clustering approaches are beginning to make their way to clinical guidelines on antidepressant selection [36], although the findings are still primarily retrospective rather than prospective.

Collecting bigger data

Given the promise of large samples and the possibility of collecting clinical measures even in large genomic biobanks (e.g. [37]), what big data would be most valuable for psychiatric research? Online services like Amazon Mechanical Turk (AMT) allow large samples to be rapidly tested on cognitive tasks, allowing robust assessment of novel task characteristics and their relationship to self-reported clinical symptoms [38]. For example, Gillan et al. tested 1413 individuals on AMT on both a cognitive task and standard questionnaires covering nine diagnostic categories [39]. Their analysis revealed three underlying trans-diagnostic

**Box 1 Explicit assessment of validation is critical**

It is likely that we are collectively deluding ourselves about the true generalizability of our models. Validation is typically assessed with leave-one-out approaches or at best k-fold cross-validation within a single study. However, both approaches are known to overestimate predictive accuracy [59], and are better interpreted as demonstrating that model performance was not grossly influenced by a small number of individuals in the training data. Leaving individual sites out of a multi-site study is a slight improvement, but is still considered internal validation [60]. For example, Koutsoulis tested a first-episode psychosis algorithm on sites left out of an international multi-site RCT and found significant site-to-site variability [10].

When validation has been explicitly tested on an independent sample, performance invariably degrades. Chekroud et al. found that the predictive accuracy of an antidepressant treatment algorithm that was 65% accurate during cross-validation was only 60% accurate in an independent RCT cohort, even though the training data included thousands of patients [34]. A large consortium found no replicable genetic result associated with treatment response in 3756 patients with major depression [62].

Medicine is riddled with observational or retrospective findings that initially seemed promising, but eventually did not improve outcomes (e.g. the idea that vitamins might prevent cardiovascular disease). As a field, we should implement strict replication and validation procedures to avoid undermining the integrity of the care we are attempting to provide. [25,65]. A bare minimum is requiring explicit assessment of validation, and a framework for setting individual results into the broader context of findings between pure research and clinical practice [57]. We should also be clearer about how any one finding relates to outcomes. Genome-wide association studies still tell us little about the treatments that will help patients get better, despite samples numbering hundreds of thousands. The currency of medicine is patients getting better and not model parameters. This fact bears repeating.

www.sciencedirect.com
factors: anxiety/depression, compulsion and intrusive thought, and social anxiety. The compulsion and intrusive thought factor selectively related to task performance, a result replicated in a separate experiment. The same trans-diagnostic factor was associated with high confidence but low self-evaluation ability while the anxiety/depression factor was associated with low confidence but high self-evaluation ability [40]. AMT has also been used to describe the landscape of emotional experience using a factor analysis on emotion labels and ratings elicited by 2185 videos [41]. Concerns have been raised about the reliability of online testing. Reassuringly, depressive symptoms on AMT have high test–retest reliability ($r = 0.87$) after one week [42], suggesting that online data can be sufficiently reliable. More directly relevant to mental health are attempts to leverage online data acquisition and testing for therapeutic means. For example, a smartphone study of digital cognitive behavioral therapy (CBT) in individuals with insomnia successfully reduced insomnia, and mediated reductions in paranoia and hallucinations [43].

Online testing provides a rapid means of prototyping new paradigms in a diverse sample of individuals with a broad range of common clinically relevant symptoms. Depression and anxiety have among the highest costs of all psychiatric disorders and are both as common in online samples as in the general population [42]. Hence, online data collection platforms may allow data to be collected quickly and cheaply, providing an opportunity to acquire enough data to support more sophisticated analyses. There are some downsides to this approach. An online community of workers might also have systematic biases that complicate interpretation of results. Inciting individuals to truthfully reveal information about their private introspective state is also difficult, although some people may be more willing to reveal personal information anonymously than in the presence of a researcher. Promising results obtained in anonymous online studies should be replicated in both more controlled laboratory studies and less controlled smartphone studies.

**Mobile devices reveal rich real-time data in the natural environment**

Particularly noteworthy are opportunities provided by mobile devices like smartphones. The data that can be collected are extremely rich, spanning ‘active’ cognitive task and self-report data with consistent results (Figure 2a) for equivalent tasks delivered in the laboratory and by smartphone [44–46]. A smartphone study in 18420 participants showed that momentary happiness reflected whether things are going better than expected, and not how well things are actually going [47]. Happiness in the computational model could be predicted by the history of rewards and expectations and related to fMRI measurements in the striatum. Depressive symptoms did not affect the strength of this relationship, but instead altered the set point around which happiness varies (Figure 2c) [48**]. Repeated sampling of emotional self-reports has also started to identify dynamic qualities in depression over months and even years. Depressed patients with worse outcomes over several years show stronger inter-connectedness of symptoms [49], and transitions between states of wellness and depression suggest that these states have features that render them inherently stable (Figure 2d) [50]. Features associated with stable healthy and depressed states might be identifiable...
in individuals [51], allowing targeted treatments when such features are detected.

Smartphones passively collect a number of streams of data that might eventually contribute toward a richer understanding of psychopathology. Elements including acceleration, movement, location, voice, keyboard use patterns, and social interactions through many sensors might eventually form part of a digital phenotyping of individuals [52,53]. For example, accelerometer data on 91,105 UK Biobank participants were used to derive circadian rhythmicity parameters related to sleep-activity patterns. Circadian disruption was associated with increased lifetime risk of both major depression and bipolar disorder [54].

Theory-driven and data-driven approaches are complementary

Data-driven analyses are agnostic about the nature of the data being analyzed. This can be a weakness because algorithms might neglect what science has already established about the structure of the data. Knowledge of the generative processes can be used to derive so-called sufficient statistics, statistically optimal summaries capturing non-random variation in the data. General-purpose machine-learning algorithms are not obliged to recover sufficient statistics, while theory-driven models are well placed to do so. Hence, theory-driven models can efficiently summarize complex data given things that we know about a generative process, and these summaries can provide inputs for machine-learning algorithms. Combining theory-driven and data-driven approaches can outperform data-driven approaches alone [6]. Figure 1b provides an example for the case of reinforcement learning. Widely used in the machine learning community, reinforcement-learning algorithms have been extensively examined in neuroscience experiments and the generative processes are increasingly understood at the neurobiological level [55,56].

How do we ensure big data and machine learning improve outcomes?

First, there is a need to generate large data sets that allow the community to parse the complexity inherent in treating psychiatric illness. We have pointed out a number of ways to generate and use large data sets. This will require collaboration, both among academics and with industry. The experience and success of the Psychiatric Genetics Consortium, along with academic–industry partnerships [37] are good examples of how fruitful this approach can be. With sufficiently large data sets, machine-learning techniques can better reach their potential for identifying more expressive functions in data, such as non-linearities or higher-order interaction effects that might not emerge readily using traditional statistical techniques.

Second, recovery of imperfect diagnostic labels is of limited value, as is comparison of cognitive and neural processes between healthy and patient populations with a specific diagnosis. The progress over the past thirty years
has shown that progress with these approaches has led to few major breakthroughs. Instead, we suggest that neurobiological research must examine interventions or risk failing to address the core aim of psychiatry, alleviating the burden of illness. Interventional research is more difficult and risky than cross-sectional research. Interventions throw up complex ethical issues and necessarily involve longitudinal designs, which are burdensome for both participants and experimenters. It is particularly risky for early career scientists because failures are costly yet common.

Clearly, these issues span different aspects in the development of novel methods. We recently suggested a roadmap inspired by the drug development pipeline [57] (Figure 3). This distinguishes early phases in which relevant probes are developed and intermediate phases in which the relationship of these probes to clinical issues (particularly treatment) are established. In later phases (akin to phase III or phase IV clinical trials), use of probes is examined with respect to clinical improvement. Adopting such a framework might help structure the field and clarify the importance of examining the relationship between probes and clinical outcomes.

Finally, we hope that big data and machine learning might spearhead a move toward clinically relevant outcomes that nevertheless allow a more direct link to particular neural or cognitive processes. Currently, outcome measures in mental health either reasonably index severity by collapsing across symptoms likely related to neurobiologically distinct domains, or reasonably index individual neurobiological processes but fail to relate to mental health in a clinically relevant manner. For example, an ICD-10 or DSM-5 diagnosis of depression requires a combination of symptoms that can include anhedonia, depressed mood, sleep disturbances, and changes in concentration. The presence of these symptoms indexes suffering, need for care, and important long-term outcomes namely mortality, remission, and return to work, but may not directly relate to any neural or cognitive process. In contrast, detailed measures of motivation can be mapped onto precise neural substrates, but variation in individual processes bears insufficient relationship to the severity of the illness, that is paramount for clinical practice.

**Dangers in the application of machine learning to psychiatry**

As a medical discipline, psychiatry’s first aim is to do no harm. The collection of useful big data in mental health requires deeply personal information. Abuse of this data is possible given the profound stigma and social impairment concomitant with mental illness. At the same time, the specific data available to machine-learning approaches can lead to biases if certain types of data are not well represented. Algorithms can inherit the biases present in the training data sets (e.g. semantics derived from standard internet-based language corpora exhibit common race and
gender biases [58]). Given that psychiatric care encompasses compulsory treatment, black box algorithms have the potential to be profoundly discriminatory. This is one issue that the European Union General Data Protection Regulation (GDPR) addresses with the requirement of appropriate procedures to prevent algorithmic biases.

Another danger is the growing perception that machine-learning algorithms are a viable solution in low-N high-p settings, that is, settings where the number of predictors exceeds the number of individuals in the sample. It is correct that machine-learning algorithms will converge in these settings, whereas traditional statistical approaches (e.g., logistic regression) will fail due to unique separation and fitted probabilities of zero or one occur. Many researchers interpret this as an excuse to apply machine-learning algorithms to small samples. However, there is no silver bullet that can replace collecting enough data to generate stable and generalizable predictions [31,34,59]. Although algorithms like the elastic net (a form of penalized regression) and tree-based ensembles (e.g., XGBoost, gradient boosting machines, or random forests) are often used and will fit in low-N high-p settings, caution should be taken to validate the performance of these models on an independent sample in line with established procedures [60,61]. If these caveats are addressed, we believe that these new approaches combining big data, machine learning and theory-driven development of probes have enormous potential to improve treatment of mental illnesses.

Conflict of interest statement

Dr. Chekroun reports holding equity in Spring Care, Inc., Fitbit Inc and UnitedHealthcare Inc. He is lead inventor on three patent submissions related to treatment for major depressive disorder (USPTO docket no. Y0087,70116US00, provisional application no. 62/491,660, and provisional application no. 62/629,041). He has consulted for Fortress Biotech on matters relating to antidepressant drug development and also holds equity in Fitbit Inc and UnitedHealthcare Inc. Dr. Huys and Dr. Rutledge report no financial or other conflicts of interest.

Acknowledgments

Dr. Rutledge acknowledges support from a Medical Research Council Career Development Award (MR/N02401X/1) and a NARSAD Young Investigator Award from the Brain & Behavior Research Foundation, P&F Fund. Dr. Huys acknowledges support from the UCLH NIHR BRC. The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


Developed and validated a personalized advantage index by deriving separate predictors from each arm of a randomized clinical trial.


Re-published a finding whereby antidepressant treatment allocation through analysis of quantitative EEG outperformed clinical decision-making.


This study used an online platform to collect data in a reinforcement-learning task from a large sample. Task performance related to a single trans-diagnostic factor estimated from questionnaires covering nine diagnostic categories.


This study used a smartphone application to capture rapid fluctuations in mood during a decision-making task on a smartphone platform. Baseline mood parameters were lower in depressed participants, linking emotional states during smartphone tasks to major depression.

49. van Borkulo C, Boschloo L, Borsboom D, Penninx BWH, Waldorp LJ, Schoevers RA: Association of symptom network
structure with the course of depression. *JAMA Psychiatry* 2015, 72:1219-1226.


Described a roadmap for the development of computational psychiatry tools for clinical practice.


