One Size Does Not Fit All:  
Methodological Considerations for Brain-Based Predictive Modelling in Psychiatry

Running Title: Methodological Considerations for Predictive Modelling

Authors:
Elvisha Dhamala¹-²,*, B.T. Thomas Yeo³⁶, Avram J. Holmes¹²⁸⁹,*

Affiliations:
¹ Department of Psychology, Yale University, New Haven, USA
² Kavli Institute for Neuroscience, Yale University, New Haven, USA
³ Centre for Sleep & Cognition & Centre for Translational Magnetic Resonance Research, Yong Loo Lin School of Medicine, Singapore, National University of Singapore, Singapore, Singapore
⁴ Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
⁵ N.1 Institute for Health & Institute for Digital Medicine, National University of Singapore, Singapore, Singapore
⁶ Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore, Singapore
⁷ Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA
⁸ Department of Psychiatry, Yale University, New Haven, USA
⁹ Wu Tsai Institute, Yale University, New Haven, USA

* Corresponding authors: Elvisha Dhamala (elvisha.dhamala@yale.edu) and Avram J. Holmes (avram.holmes@yale.edu)

Keywords: machine learning, predictive modelling, brain, neuroimaging, biomarkers, algorithms
Abstract
Psychiatric illnesses are heterogeneous in nature. No illness manifests in the same way across individuals, and no two patients with a given diagnosis will exhibit identical symptom profiles. Over the last several decades, group-level analyses of in vivo neuroimaging data have led to fundamental advances in our understanding of the neurobiology of psychiatric illnesses. More recently, access to computational resources and large publicly available datasets alongside the rise of predictive modelling and precision medicine approaches have facilitated the study of psychiatric illnesses at an individual-level. Data-driven machine learning analyses can be applied to identify disease-relevant biological subtypes, predict individual symptom profiles, and recommend personalized therapeutic interventions. However, when developing these predictive models, methodological choices must be carefully considered to ensure accurate, robust, and interpretable results. Choices pertaining to algorithms, neuroimaging modalities and states, data transformation, phenotypes, parcellations, sample sizes, and populations we are specifically studying can influence model performance. Here, we review applications of neuroimaging-based machine learning models to study psychiatric illnesses and discuss the effects of different methodological choices on model performance. An understanding of these effects is crucial for the proper implementation of predictive models in psychiatry and will facilitate more accurate diagnoses, prognoses, and therapeutics.
Introduction
Progress across the brain sciences has provided unprecedented opportunities for advancing our understanding of the etiology and pathogenesis of psychiatric illness. In the last several decades, new *in vivo* imaging techniques, such as magnetic resonance imaging, have facilitated the successful identification of neural correlates of human behavior. Traditionally, patient populations were treated as distinct entities, and corresponding group-level analyses shed light on the neurobiological changes that characterize these nominally distinct diagnostic constructs. However, the heterogeneity in psychiatric illnesses and murky boundaries between health and disease make it difficult to establish the extent to which brain-behavior relationships are consistently or uniquely disrupted. To address this heterogeneous nature and circumvent diagnostic biases, precision medicine has emerged as a powerful tool to understand clinical presentation at an individual-level by considering a range of biological and environmental factors(1). We are currently at an exciting time where we can harness the joint powers of precision medicine, machine learning, and neuroimaging to enhance our understanding of the neurobiology of psychiatric illnesses and associated symptom profiles. Increased access to computational resources and large publicly available datasets are supporting a field-wide shift to data-driven discovery science, allowing descriptive models of psychiatric illness, predictions of clinically-relevant behaviors, and personalized prescriptions for therapeutic interventions ushering in a new era for psychiatry.

In psychiatry, precision medicine and predictive modeling have two primary goals: to discover how the brain gives rise to behavior and to identify practically useful markers of behavior and clinical outcome. Precision medicine transitions away from the *one-size-fits-all* model of traditional medicine and seeks to identify the underlying causes of disease and appropriate therapeutic interventions at a patient-specific level. This individualized approach has led to significant advances across the health sciences, for instance in oncology where personalized treatments are now available for a range of cancers(2). The application of predictive models in psychiatry, termed precision psychiatry(3), incorporates insights from molecular biology, cognitive neuroscience, neurophysiology, and behavioral and clinical psychology to capture clinically-relevant biomarkers. If successful, precision psychiatry will transform how we diagnose and treat psychiatric illnesses.

There is growing interest in the development and application of novel tools to predict illness risk, disease progression, and identify shared and unique properties of diagnostic groups and
associated symptom profiles. Psychiatric diagnoses categorize clusters of co-occurring symptoms that often lack clear discernible borders between health and disease (4). While the presence of specific symptom profiles can distinguish diagnoses, phenotypic presentations can overlap across patient groups making it difficult to identify specific markers of illness. In this regard, discrete and dimensional models of psychiatric illness provide complementary information. Discrete medical models provide diagnostic labels based on clinical descriptions of signs and symptoms, while dimensional models describe broad hierarchical features of psychopathology. Together, different models of psychiatric illness can facilitate improved understanding of psychopathology at an individual-level.

Traditional analyses in psychiatry have largely sought to uncover group-level patterns characterizing distinct illnesses. While these analyses have provided foundational insights in our understanding of psychiatric illness, they fail to adequately capture individual variability in symptom presentations. Machine learning provides a rich analytic framework to reveal the neurobiological correlates of discrete and dimensional measures of psychopathology at an individual-level. By leveraging our understanding of brain-behavior relationships and considering a specific patient’s personalized profile, machine learning can be used to generate individual-level behavioral predictions. Machine learning analyses can be characterized into three groups: descriptive, predictive, and prescriptive, each with distinct utility for psychiatry researchers (Figure 1). These analyses rely on supervised, semi-supervised, and unsupervised algorithms. As an exhaustive overview of different analyses and algorithms is not possible here, we provide key terms and examples in Table 1 (for additional discussion see(5)).

Guidelines for implementing predictive models to study behavioral variability based on neuroimaging data have been proposed (5-11), but there has not yet been a review of how different methodological considerations can influence the models. Here, we review the effects of algorithm selection, neuroimaging modality, data transformation, sample sizes, brain parcellation, phenotype selection, and population-specific approaches on predictive models and contextualize how insights gained from those analyses can facilitate advances in psychiatry (Figure 2).

**Algorithm Selection**
The structure of neuroimaging data presents a unique set of analytic challenges in terms of dimensionality, collinearity, and sparsity. To address these challenges, different algorithms have
been designed including the connectome-based predictive model(12) and BrainNet convolutional neural network(13). The benefits of these cutting-edge approaches include: (1) data-driven feature selection algorithms that do not rely on brain regions or connections selected a priori, but rather account for the fact that behavior and cognition may be best characterized as properties of covariation across large-scale networks; (2) built-in cross-validation, which ensures models are generalizable and replicable; and (3) the potential for continuous predictions that, in contrast to categorical or diagnostic predictions, provide more dimensional measures of psychopathology. Connectome-based predictive modelling is a data-driven protocol to predict behavior from connectivity data. Here, researchers determine which connectivity features are most strongly correlated with a given phenotype, and then use those features to build a linear model to predict behavior. This approach has identified functional connectivity correlates of attention(14, 15), memory(16), and anxiety(17) across distinct populations, and predicted responses to antidepressants(18) and treatment for cocaine and opioid use disorders(19, 20). BrainNet relies on a convolutional neural network framework that maintains the topological complexity of brain networks. This framework has been applied to predict future cognitive and motor development in preterm infants based on white matter connectivity(13).

While both connectome-based predictive models and BrainNet convolutional networks support the analysis of connectomic data, other classical machine learning and deep learning algorithms can also generate meaningful predictions. Amidst ongoing debates on which algorithms yield the most accurate predictions, head-to-head comparisons suggest a suite of approaches may be clinically useful. In one recent comparison, kernel ridge regression, a generic feedforward network, a BrainNet convolutional network, and a graph convolutional network generated comparable cognitive predictions based on functional connectivity(21). A separate analysis also reported that classical approaches, including elastic net (which optimizes between L1 and L2 regularization), achieve comparable or better predictive accuracies than deep learning methods to capture associations between functional connectivity and cognition(22). However, findings have been mixed, and a separate comparison revealed deep learning models produced the best cognitive predictions based on whole brain gray matter volume in healthy and clinical (mild cognitive impairment, Alzheimer's disease) populations(23).

Comparisons of classical and deep learning approaches for classification problems have also produced mixed findings. One study found deep learning classifiers were more accurate when
relying on anatomical or functional data (or both) to distinguish individuals with autism from healthy controls (24). On the contrary, an international challenge to predict autism spectrum diagnosis from anatomical and functional data revealed deep learning models were more likely to overfit such that they failed to generalize to different datasets whereas simpler models achieved more robust performance (25). Relatedly, across different classical algorithms, non-sparse linear methods more accurately distinguish different psychiatric diagnostic groups from healthy controls than sparse linear or non-linear methods (26).

Hence, different machine learning algorithms are suitable for distinct predictions, and algorithm selection should consider whether accuracy, interpretability, or both should be prioritized and be primarily driven by the goals of the study.

**Neuroimaging Modality**

Different neuroimaging modalities capture distinct information about the underlying neurobiology. Structural MRI provides information about the shape, size, and integrity of gray and white matter brain structures, and can be used to study physical damage and lesions. Functional MRI indirectly measures brain activation by detecting changes in blood oxygenation, facilitating the study of functional activation patterns and correlated signals across functional systems. Finally, diffusion MRI measures diffusion of water molecules in biological tissues and enables the quantification of white matter tracts throughout the brain. These modalities yield distinct biologically relevant features for the study of psychiatric illnesses and prediction of clinically-relevant phenotypes.

Converging evidence suggests that individual differences in cognition, personality, and emotional traits, as well as psychiatric diagnoses, are most accurately predicted by functional connectivity (24, 25, 27-31). These results, reported across healthy and clinical populations, suggest that functional MRI data may be the best modality for predictive modelling. However, there is evidence to suggest that other modalities may also produce clinically informative predictions, especially in psychiatric illnesses associated with structural abnormalities in the brain.

Cortical gray matter volume and thickness successfully predicts diagnosis of Alzheimer’s, autism, and attention deficit disorders (32) and symptom burden in schizophrenia (33). Gray matter density reliably estimates response to repetitive transcranial magnetic stimulation in
patients with schizophrenia(34). Finally, voxel-level structural scans distinguish whether a patient with depression may be treatment-resistant(35).

Meanwhile, diffusion-derived measures of fractional anisotropy along with radial, mean, and axial diffusivity predict response to cognitive behavioral therapy in social anxiety disorder(36). White matter connectivity characterizes tic improvement following deep brain stimulation for Tourette syndrome(37) and distinguishes healthy controls from patients with mild cognitive impairments and patients with dementia(38).

**Intrinsic vs. Task-Evoked Measures of Brain Functioning**

Intrinsic (or resting state) functional connectivity captures the extent to which neural activation patterns between a pair of brain regions covaries over time, providing insight into general functional network organization in the brain. Task-evoked functional connectivity characterizes covariations in neural activation patterns while performing specific tasks and facilitates the understanding of the functional networks that underlie different behaviors. While both intrinsic and task-evoked measures of functional network connectivity can generate insightful behavioral predictions, models based on task-evoked connectivity more successfully predict individual cognitive traits(39, 40). Moreover, stacked models incorporating multimodal data yield the greatest accuracy(29). Intrinsic striatal connectivity predicts responses to antipsychotic drug treatment in patients with schizophrenia and acute psychosis(41), while whole-brain intrinsic functional connectivity distinguishes bipolar and major depressive disorder(42), and estimates risk of developing anxiety disorders(17). Meanwhile, connections evoked during emotional face perception tasks meaningfully distinguish individuals with social anxiety disorder, panic disorder, and healthy controls(43). Task-evoked connectivity using memory and general executive function tasks also transdiagnostically predicts memory deficits across schizophrenic, bipolar, and attention deficit hyperactivity disorders, and these models generalize to healthy participants(16). Alternatively, task-evoked measures of brain activation are also clinically meaningful. Activation patterns discriminate between individuals with schizophrenia and bipolar disorder(44), distinguish between major depression with and without mania from bipolar depression(45), and predict clinical responses to cognitive behavioral therapy in social anxiety disorder(46) and in panic disorder(47).

The acquisition of high-quality functional MRI scans sensitive to motion may be difficult to achieve in psychiatric populations. Patient wellbeing can be prioritized while addressing this to
acquire high-quality data using different strategies including familiarizing subjects with the scanner environment, using physical cushioning/restraints, and implementing motion correction strategies (48-51). Moreover, while task scans predict behavioral traits more accurately in healthy populations, they can be more cognitively demanding on participants, require additional scan time, and present language barriers, making them less feasible in clinical samples. To circumvent this issue, resting-state data can be used to synthesize task data (52) and naturalistic stimuli (movie-watching) data (53). Those synthesized data can predict individual differences in brain activity during task performance (52), naturalistic viewing (53), and behavioral traits (54) in healthy populations. In a clinical context, this approach can facilitate more accurate behavioral predictions in a psychiatric population, reducing (and even eliminating) the monetary and cognitive resources associated with the acquisition of task and naturalistic scans.

Data Transformation

To transform noisy estimates of brain structure and function into reliable predictors of human behavior, neuroimaging data must be carefully preprocessed and cleaned. Preprocessing pipelines and analytical methods can introduce variability into neuroimaging studies and subsequently bias findings (55-58).

Before any processing, analyses, or modelling can be performed, neuroimaging data must undergo quality control to ensure images are free of artifacts and have a high signal-to-noise ratio. Unfortunately, stringent quality control is likely to result in samples that are less representative of the population in terms of age, sex, race, and socioeconomic status and remove meaningful variance in the data (40). Lower data quality may also be indicative of and associated with psychiatric illness. Hence, careful consideration must be given to arbitrary quantitative and qualitative data quality thresholds that may bias subsequent predictive modelling analyses.

Relatedly, missing imaging and behavioral data is also a particular point of concern in clinical populations. Missing data may be a direct result of data entry errors, participant difficulty in completing data collection, attrition at follow-up, or an unintended consequence of stringent quality control (59). Similar to quality control, missing data is more likely to be present in clinical populations, especially those suffering from severe psychiatric illness, and individuals from specific demographic groups that are likely already underrepresented. Although potential solutions for missing data imputation have been proposed, the field has not yet reached a
consensus about best practices. Future research should incorporate lessons learned from decades of clinically-oriented research seeking to capture complex clinical outcomes using incomplete data.

Global signal regression is a highly debated method used to remove the effects of global variations in activity from the functional time series of each voxel(60, 61). Most, although not all, behavioral predictions across cognitive, personality, and emotional domains based on intrinsic and task-based functional connectivity benefit from global signal regression(39, 62).

Another critical analytical choice for functional MRI addresses the parametrizations of the functional connectome. Full correlation is a standardized form of covariance. Partial correlation is derived using the inverse of the precision matrix and represents the correlation between a pair of brain parcels after adjusting for the time series of all other parcels. Furthermore, not every matrix is a valid correlation (or covariance) matrix. Tangent space representations allow the manipulation of the functional connectivity data consistent with the mathematical properties of correlation (or covariance) matrices(22). Some studies have suggested that compared with full and partial correlations, tangent space representations yield more accurate and reliable diagnostic classifications across distinct psychiatric illnesses(26).

Relatedly, white matter connectivity is defined using deterministic and probabilistic tractography algorithms(63, 64). Deterministic tractography relies on fixed orientations to direct streamlines while probabilistic algorithms estimate a distribution of fiber orientations and randomly draw a sample from this distribution to direct streamline propagation(64). White matter connectivity derived using probabilistic tractography more accurately and reliably predict cognition than deterministic tractography(27). Taken together, these results suggest that connectomic definitions can also bias behavioral predictions.

Additionally, the use of anatomical features for predictive modelling involves crucial data transformation steps that can bias behavioral predictions. Individual differences in intracranial volume are typically considered to be confounds(65) but proportional correction of those individual variations can differentially bias cognitive predictions based on cortical surface area, gray matter volume, and thickness(66). To better capture clinical deviations in brain structure and function, normative modelling can be used to model expected variations in a given neuroimaging feature(67). Deviations from normative models of regional cortical volumes more
accurately predict overall and dimension-specific psychopathology compared to raw measures of cortical volume (68).

Critically, global brain activation (69), intracranial volume (70, 71), regional cortical volume (72-75) have been implicated in and may be indicative of psychiatric illness. Hence, it is crucial to account for the underlying, clinically-relevant signals that may be captured in distinct properties and consider how they should be handled when developing predictive models to study psychiatric populations.

**Phenotype Selection**

An unresolved debate pertains to the structure of psychiatric illnesses and whether researchers should focus on diagnoses, symptoms, or dimensional profiles (76-78). Psychiatric diagnoses rely on a clinician’s description of the syndrome’s nature, timespan, severity, and pre-existing risk factors to establish a categorical classification (79). These descriptions may be biased and the resulting diagnoses unreliable. In fact, two of the most common psychiatric illnesses, major depressive disorder and generalized anxiety disorder, have inter-rater reliabilities of 0.28 and 0.21, respectively (80). As a result, clinically distinct phenotypic profiles can be assigned the same diagnosis. Remarkably, there are 636,120 distinct clinical presentations of post-traumatic stress disorder (81). It is naïve to believe the same neurobiological properties will underlie the unique clinical presentations of a given diagnosis (82). Further impacting the clinical utility of brain-behavior models, clinical scales can vary substantially in terms of the symptoms and dimensions they capture, and single symptomatic or dimensional measures are insufficient to capture the complex nature and presentation of a psychiatric illness in an individual. These data raise many questions about which phenotypes or diagnoses we should aim to predict and whether those predictions will be biologically meaningful. Data scientists are familiar with the idea of “garbage in, garbage out”: if relying on poor phenotypic data and/or labels to train predictive models, the predictions and underlying associations captured are also likely of poor quality. Therefore, to maximize the accuracy and interpretability of prediction models in psychiatry, the variables being predicted must be reliable, validated, and representative of the overarching psychiatric illness or phenotype being studied in the specific population of interest.

Directly addressing this pitfall, researchers have begun to develop frameworks to reconceptualize diagnostic criterion of psychiatric illnesses into quantifiable behavioral measures, such as the Extended Strengths and Weaknesses Assessment of Normal
Behavior(83). Moreover, the use of passive digital data collection (i.e., from smartphones and wearable technology) along with better-distributed self-report measures is likely to yield more representative and robust quantifications of clinical behaviors. Overall, an improvement on the quantification of complex behaviors is crucial for the successful implementation of predictive models to capture relationships between those behaviors and neurobiological markers.

Ongoing research also suggests certain illnesses, behaviors, or phenotypes may be more predictable than others. Intrinsic and task-based functional connectivity, and structural properties yield higher accuracies when predicting cognitive traits than personality or mental health traits(40, 84). Within the cognitive domain, composite scores summarizing broad cognitive domains (85-87), are more accurately predicted than individual task scores(88). Therefore, in addition to a variable being valid, reliable, and of clinical interest, its associations with neurobiological properties (and the underlying signal to noise ratio) must also be strong enough to be captured using multivariate approaches.

Relatedly, evaluation of model generalizability across datasets can be tricky as they may collect different clinical scales and psychometric measures. Fortunately, recent attempts to characterize patient profiles using symptomatology and behavioral measures revealed that specific connectivity patterns underlie multiple phenotypic expressions(89). While further research is needed in this area, these findings suggest that, in some instances, a single model may be generalizable across datasets or populations to predict different independent, but phenotypically correlated, measures.

**Brain Parcellations**

Machine learning analyses using neuroimaging data are prone to the curse of dimensionality. Imaging data at a voxel-level contains hundreds of thousands of features. Brain parcellations reduce dimensionality by segmenting the brain into tens or hundreds of regions or networks(90, 91). These parcellations may be defined based on cytoarchitectural organization(92-94), sulcogyrical localization(95-97), data-driven modelling of intrinsic(98-103) or task-evoked(104) functional activation signals, or independent component analyses(105, 106). Dimensionality reduction through parcellation reduces noise, decreases computational complexity, and enhances interpretability(107), but can also result in loss of biologically meaningful signals captured at a voxel-level and obscure valuable individual-level variations in brain organization(107-109). While some analyses have shown evidence that parcellation resolution
and definition can affect predictive performance(26, 27, 110, 111), others have reported generally consistent results across parcellations(24).

To eliminate parcellation biases, ensemble approaches integrating information across parcellations have been proposed(112, 113). Ensembling across pre-defined parcellations improved the classification of individuals with schizophrenia versus healthy controls based on intrinsic functional connectivity and regional activation properties compared to predictions based on a single parcellation(113). Similarly, ensembling across pre-defined or stochastic parcellations yielded more accurate and robust classification of individuals with autism versus healthy controls based on intrinsic functional connectivity than predictions based on a single pre-defined parcellation(112).

A second alternative involves individual-specific parcellations. These parcellations consider individual differences in spatial arrangement of the brain and are likely to be critical in psychiatric populations with known differences in network topography(107, 109, 114, 115). These individual-specific network topography patterns have demonstrated improved accuracy for behavioral predictions compared to population-level parcellations(107).

**Sample Sizes and Transfer Learning**

Historically, psychiatric neuroimaging studies relied on as few as tens of subjects to characterize interactions between neurobiology and diagnoses/phenotypic presentations(116). A core issue when applying machine learning to neuroimaging is sample sizes. Generally, machine learning algorithms require sample sizes larger than the dimensionality of the feature set to prevent overfitting(117). Over the last decade, sample sizes have increased to hundreds and even thousands. Limited sample sizes result in decreased reliability of predictions regardless of algorithm, feature type, and phenotype(84, 118-120), and produce overly optimistic estimates of prediction accuracy(117, 121). These issues are of particular concern in heterogeneous populations, such as individuals with psychiatric illnesses, where small datasets will likely produce models that cannot generalize to independent test samples and thus lack clinical utility. In fact, there appears to be a strong negative relationship between accuracy and sample size for diagnostic and behavioral predictions in Alzheimer’s, depression, schizophrenia, psychosis, and autism(117, 120). However, large clinical samples can be difficult to acquire.
Fortunately, there is evidence to suggest that brain-behavior predictive models developed in the general population can boost predictions of new phenotypes in smaller independent datasets. A “meta-matching” framework developed to translate predictive models based on neuroimaging features from large samples into non-brain imaging features in smaller samples has shown promising results when generalizing from aging to young adult populations(122). Connectome-based predictive models of attention have also demonstrated the capacity to generalize across healthy and clinical populations(14, 15, 123), while models of memory have successfully generalized across psychiatric illnesses(16). Translational applications of these predictive frameworks provide the opportunity to leverage information about brain-behavior relationships in the general population or within a specific clinical population to better understand psychiatric presentation. This can drastically reduce the clinical sample size needed to yield accurate and reliable predictions.

**Population-Specific Models**

The rise of machine learning applications in neuroscience and psychiatry has begun to expose how (lack of) demographic representation in datasets and algorithmic biases can inadvertently introduce biases in behavioral predictions for specific populations. These discrepancies cannot be entirely attributed to under-representation in datasets or unequal sample sizes and may also be tied to differences in how accurately and reliably neuroimaging and behavioral measures are able to capture phenotypes across populations(124). Evidence suggests that behavioral prediction models trained on datasets dominated by White Americans may fail to generalize in African American populations and the underlying brain-behavior associations captured by the models are specifically representative of White Americans(125). Similarly, predictions of cognitive and personality traits can fail to generalize across sexes(88, 126-128). Distinct demographic groups (i.e., age groups, sexes, races/ethnicities) often exhibit differences in psychiatric presentations including prevalence, symptom profile, and treatment response. Traditional neuroimaging analyses relied on covariate regression to minimize effects of sex, age, and/or other demographic or data collection dependent variables (i.e., site). However, common methods to remove covariate information may not sufficiently account for their effects(66, 129) and are likely to result in a loss of meaningful variance in the data. Covariate regression, if not performed properly within nested designs, can also lead to information leakage, and artifically inflate model performance and effect sizes. Alternatively, the use of population-specific prediction models can provide insight into how brain-behavior relationships vary across different populations.
While general predictions can capture shared neural correlates of different clinically-relevant behaviors, population-specific models can provide complementary information about brain-behavior relationships specific to different groups. Recent work has shown that general prediction models often fail in individuals who do not fit a sample population’s stereotypical profile for brain-behavior relationships (130). Demonstrating the value of population-specific predictive modelling, several groups have shown that sex-specific predictions of cognition and personality are more accurate and insightful than sex-independent ones (39, 126-128). Specifically, these models reveal the shared and unique correlates of behaviors in males and females. Together, these data emphasize the clinical utility of population-specific predictive models. Given the well-established sex differences in prevalence, symptomatology, and clinical trajectory in psychiatric illnesses, sex-specific predictive models will be especially informative in distinct clinical populations. Similarly, precision psychiatry will likely also benefit from models specific to other demographic and clinical variables that can drive differences in disease presentation and treatment seeking behaviors.

A crucial factor to acknowledge when developing population-specific models is intersectionality. The overlapping and interdependent nature of how distinct demographic and social identities (i.e., age, sex, race, ethnicity, sexuality, physical health, psychiatric diagnoses, socioeconomic status, religion) may intersect is extremely complex and individuals suffering from psychiatric illnesses may be members of several different marginalized groups (131). Other variables pertaining to data collection (i.e., data quality, site) are likely to also be closely intertwined with participants’ intersectional identities. Hence, when developing population-specific models, researchers must carefully consider how models may generalize across intersectional populations. Models specific to one population may not comparably generalize across other demographic and social groups, but they may still reveal brain-behavior associations specific to that group that can aid in clinical decision-making. Moreover, ensembling predictions across models developed across distinct populations that an individual belongs to may also provide clinically-relevant insights.

**Clinical Utility and Future Directions**

An ongoing debate in the field seeks to establish whether predictive models should prioritize accuracy, generalizability, or interpretability. Existing literature suggests a wide range of modelling choices can be leveraged to yield clinically-informative predictions. Models with high
accuracy and interpretability but low generalizability might offer important insights into brain-behavior relationships in clinical populations but lack clinical utility. Conversely, models with high accuracy and generalizability but low interpretability may have high clinical utility but are unable to provide clear explanations that can aid clinical decision-making. Finally, while models with high interpretability and generalizability but low accuracy are extremely unlikely by definition, they can produce mechanistic insights into the neurobiological features underlying a phenotype. Moving forward, the field of precision psychiatry must incorporate interdisciplinary researchers and carefully consider how methodological choices can be used to develop machine learning models that are in line with their broader analytical goals.

Brain-based predictive modelling in psychiatry and neuroscience has led to monumental advances in our understanding of the neurobiological correlates of psychiatric illness. In the next decade, we expect continued growth in this area of research. Two particular areas of focus crucial for significant advances include scale development and clinical evaluation. First, the development and implementation of behavioral scales that accurately, reliably, and robustly capture individual differences across distinct clinically-relevant behaviors is fundamental. Brain-based behavioral prediction models can only be successful if the behaviors they are optimized to capture are correctly measured. Second, predictive and prescriptive analyses must be prospectively clinically evaluated in appropriately pre-registered randomized controlled clinical trials. A clinical evaluation of these approaches is necessary before we can actualize and implement predictive models in psychiatry.

Conclusion
Historical advances in our understanding of neurobiological correlates of psychiatry have paralleled technological advances enabling the visualization and quantification of human brain functions. Future advances can exploit multivariate analytical tools to model accurate and interpretable relationships between brain and behavior at an individual-level. Machine learning with neuroimaging data provides a framework that can support innovations in precision neuropsychiatry. However, the development of predictive models to facilitate accurate diagnosis, informative prognosis, and impactful treatment must consider methodological choices that can influence model performance. Here, we review how different factors can affect both accuracy and interpretability of these models. An understanding of these effects is crucial for the mindful implementation of choices that are most optimal for a specific research goal. Proper development and deployment of these approaches will yield significant improvements in our
understanding of psychiatric illness and improve the lives of millions of individuals suffering from them.
Acknowledgements
This work was supported by the National Institute of Mental Health (R01MH120080 and R01MH123245 to A.J.H.) and the Kavli Institute for Neuroscience at Yale University (Postdoctoral Fellowship for Academic Diversity to E.D.). This work was also supported by the following awards to B.T.T.Y.: the Singapore National Research Foundation (NRF) Fellowship (Class of 2017), the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), and the NMRC STaR (STaR20nov-0003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore NRF or the Singapore NMRC.

Financial Disclosures
Drs. Dhamala, Yeo, and Holmes reported no biomedical financial interests or potential conflicts of interest.
References


Figure Legends

Figure 1: **Predictive models based on neuroimaging data can be applied in psychiatry to perform descriptive, predictive, and prescriptive analyses.** Information obtained from structural, functional, and diffusion scans at a population-level can be leveraged to gain insights at an individual-level. Descriptive analyses detect underlying patterns in the neurobiological data at a population-level and describe how those patterns manifest at an individual-level. These approaches rely largely on unsupervised algorithms and are technically postdictive, in that they retrospectively detect trends in existing data, rather than prospective (i.e., future) predictions. Predictive analyses capture brain-behavior relationships at a population-level and use those to predict psychiatric diagnosis, dimensions, cognition, and personality at an individual. These predictive models rely on supervised algorithms and have the potential to identify future risks for development or deterioration of illness in a personalized manner. Prescriptive analyses establish neurobiological predictors of treatment response trends at a population-level and provide recommendations on interventions at an individual-level. These approaches rely on supervised algorithms to identify anomalies and suggest personalized interventions that can most effectively alleviate individual specific patterns of concerns and/or other relevant factors. Together, descriptive, predictive, and prescriptive analyses can yield clinically meaningful insights that can facilitate the development of personalized diagnostic tools and therapeutic interventions.

Figure 2: **Methodological choices for machine learning models based on neuroimaging can bias the performance of predictive models.** Sample size: the number of individuals and the heterogeneity within the population. Neuroimaging modality: the choice of structural, functional, or diffusion imaging scans. State: functional scans may be acquired to measure activation patterns corresponding to intrinsic (resting-state), task-evoked, or naturalistic states (e.g., movie watching). Parcellations: distinct parcellations derived using structural, functional, clustering, gradient, or other information may be applied to the neuroimaging data. Data transformation: neuroimaging data must be preprocessed and transformed before it can be input into a predictive algorithm. Population-specific vs. general models: the entire sample may be considered as a single population or separated into distinct sub-samples based on different factors such as demographics or diagnosis. Algorithm: predictive algorithms include linear or non-linear traditional machine learning and deep learning approaches. Phenotype: the specific variable to be predicted may be a categorical diagnosis, dimensional measures of...
psychopathology, cognitive abilities, or personality traits. Predictive modelling in psychiatry must be implemented with careful consideration of how these factors may potentially influence model accuracy and interpretability.
<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Definition/Description</th>
<th>Examples and Additional Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyses</td>
<td>Descriptive</td>
<td>Rely on existing observations to identify underlying patterns that can describe and summarize the data in an interpretable manner</td>
<td>Identification of distinct subtypes of autism (132, 133), depression (134, 135), psychosis (136), schizophrenia (33, 137), and bipolar and schizoaffective disorders (138)</td>
</tr>
<tr>
<td></td>
<td>Predictive</td>
<td>Establish associations between independent and dependent variables in existing data, uses learned associations to predict the trends and properties in unobserved samples</td>
<td>Prediction of clinical, cognitive, and personality traits across different psychiatric illnesses (30, 139-142)</td>
</tr>
<tr>
<td></td>
<td>Prescriptive</td>
<td>Identify patterns and trends in the underlying data and recommend personalized interventions that can most effectively alleviate individual specific patterns of concern</td>
<td>Prediction of clinical responses to medications (18, 41) and cognitive behavioral therapy (36, 46, 47)</td>
</tr>
<tr>
<td>Algoritms</td>
<td>Supervised</td>
<td>Rely on pre-observed ground truth labels to learn associations between the input and output data</td>
<td>Prediction of treatment response in social anxiety disorder (46)</td>
</tr>
<tr>
<td></td>
<td>Semi-Supervised</td>
<td>Combine a small amount of labeled data and a large amount of unlabeled data to identify associations between the input and output data</td>
<td>Identification of subtypes of internalizing disorders in youth (143)</td>
</tr>
<tr>
<td></td>
<td>Unsupervised</td>
<td>Group and interpret patterns in unlabeled data to characterize data</td>
<td>Investigation of overlapping and specific correlates of common psychiatric illnesses (144)</td>
</tr>
<tr>
<td>Supervised Algorithms</td>
<td>Classification</td>
<td>Establish relationships between the input variables and discrete binary or multi-class categorical labels</td>
<td>Distinguishing healthy and clinical populations See (145) for a review on the application of clinically informative classifiers</td>
</tr>
<tr>
<td></td>
<td>Regression</td>
<td>Predict continuous variables based on the input data</td>
<td>Prediction of disease severity See (146) for a review on the use of regression algorithms to elucidate brain-behavior relationships</td>
</tr>
<tr>
<td>Unsupervised Algorithms</td>
<td>Clustering</td>
<td>Enable the characterization of data into distinct groups (clusters) based on shared underlying patterns</td>
<td>Biologically driven disease subtype identification</td>
</tr>
<tr>
<td>Models</td>
<td>Classical Machine Learning</td>
<td>Relatively 'simple' models that rely on linear or non-linear and sparse or non-sparse approaches to</td>
<td>Linear regression, logistic regression, support vector machines</td>
</tr>
</tbody>
</table>
### Deep Learning
- Subset of machine learning models that rely on one or more neural networks (e.g., perceptrons) to progressively extract higher-level features from the input data to make predictions about the output.

### Variables
- **Binary**: Discrete data that can fall in one of two categories. Presence or absence of an illness.
- **Categorical**: Discrete data that can fall in one of several categories. Subtypes of an illness.
- **Continuous**: Data can take on any value within a continuous range. Dimensional measures of impairment.

### Regularization
- **L1 (Sparse)**: Uses the absolute value of magnitude of the parameters as the penalty term, thus encouraging sparsity by shrinking least important coefficients to zero.
- **L2 (Non-Sparse)**: Uses the squared magnitude of the parameters as the penalty term, thus reducing the magnitudes of parameters.

### Model Properties
- **Complexity**: A more complex model has the capacity for more complicated relationship among the features and target variables.
- **Transparency**: Ability to understand and interpret the brain-behavior relationships captured by the model.

### Model Performance
- **Accuracy**: Correlation (or correspondence) between true and predicted values when evaluating a model on a (held out) test set that is from the same population (or dataset) as the training set.
- **Generalizability**: Correlation (or correspondence) between true and predicted values when evaluating a model on a population (or dataset) that is unique from the population in which it was trained.