Psychiatric disorders are, without exception, highly heterogeneous in terms of symptoms, disease course, and biological underpinnings. Diagnoses are made on the basis of symptoms, while at the level of the underlying biology their causes are complex and multifaceted. This becomes acutely problematic in psychiatry because biological tests to assist diagnosis or predict outcome have not been developed (1). Diagnostic categories therefore often do not map cleanly onto either biology or outcome, which forms a major barrier to understanding disease mechanisms and developing more effective treatments.

A recognition of the imperfections of psychiatric nosology is not new; the debate between “lumpers” and “splitters” (2) over the number and validity of diagnostic classifications has continued unabated for more than a century following the classifications of dementia praecox and schizophrenia proposed by Kraepelin and Bleuler (3,4). Reflecting this ongoing debate, classifications are revised with every new edition of diagnostic manuals (5,6). Data-driven approaches to address heterogeneity in psychiatric disorders have also been applied for decades, in which the dominant approach has been to partition clinical groups into more homogeneous subgroups using data clustering methods—early examples can be seen in Paykel (7) and Farmer et al. (6). These approaches have recently received renewed interest for three reasons: 1) the advent of technologies for measuring many aspects of biology noninvasively and in vivo, particularly neuroimaging and genetics; 2) advances in statistical and machine learning data analytic approaches that make it possible to extract information from complex and high-dimensional data; and 3) increasing emphasis on using biological data to tailor treatments to the needs of individual patients (“precision medicine”) (6,10). Most notably, recent funding initiatives, such as the National Institute of Mental Health Research Domain Criteria [RDoC (11)] and the European Roadmap for Mental Health Research [ROAMER (12)], have encouraged researchers to think beyond the classical case-control approach—where participants are either “patients” or “controls” based on fixed diagnostic criteria—and instead link cognitive dimensions with underlying biology while cutting across diagnostic classifications. The
hope is that this will lead to biologically grounded understanding of disease entities and ultimately to more effective, personalized treatments.

These initiatives have stimulated an increasing number of studies that have used data-driven methods to stratify many disorders, including schizophrenia, major depression, attention-deficit/hyperactivity disorder (ADHD), and autism based on many types of data, including symptoms, neuro-psychologic scores, and neuroimaging measures (13–21). We selectively review this burgeoning literature.1 We first present a didactic introduction to the most prevalent methodologic approaches for stratifying psychiatric disorders, highlighting the (often implicit) assumptions they entail. We then present an illustrative overview of studies that have used these methods to parse the heterogeneity underlying psychiatric disorders. We identify problems with current approaches and discuss the importance of validation to ensure reproducibility and ensure that clusters map onto clinically meaningful variation. We discuss emerging techniques, such as normative modeling (22), that provide means to parse heterogeneity in clinical cohorts without needing to make strong assumptions about clinical groups and evaluate the suitability of each method for meeting the objectives of recent research initiatives. Finally, we propose future developments that may help to parse heterogeneity more effectively.

METHODOLOGIC APPROACHES FOR STRATIFYING CLINICAL POPULATIONS

The predominant approach has been to subdivide clinical cohorts using statistical or machine learning methods, largely of two main types: clustering (23) and finite mixture models (FMMs) (24–26). Both are unsupervised in that they do not have access to class labels (e.g., diagnostic labels) and must find subgroups automatically based on structure within the data and heuristics used by each algorithm. In contrast, supervised methods are provided with labels that indicate the class to which each subject belongs (e.g., “patient” or “control”). Supervised learning has been successful for predicting diagnosis or outcome from neuroimaging data in research settings (27–29) but is fundamentally limited by the quality of the clinical labels and the heterogeneity within disease cohorts (29) and cannot, by definition, inform on the validity of the labels. Therefore, unsupervised methods have been more widely used for discovering latent structure within clinical groups. We present a brief introduction to clustering and FMM methods below; additional details and a didactic introduction are provided in the Supplement.

Clusteri

Clustering

The classical case-control approach can itself be phrased in terms of defining clusters and associated decision boundaries. For example, Fisher’s linear discriminant (23) uses the class-dependent mean response (e.g., in patients vs. controls) and thereby clusters the entire cohort along a decision boundary defined by the mean and class-specific covariances. More generally, given a set of data points (e.g., clinical or neuroimaging measures), clustering algorithms aim to partition the data into a specified number (K) of clusters such that the samples in each cluster are more similar to one another than to those in the other clusters. This entails defining a measure of similarity or distance between data points. One of the simplest and most widely used approaches is k-means clustering, which partitions the input space into K subregions based on the squared Euclidean distance (see Supplement). A wide variety of other algorithms have also been proposed in the machine learning literature (23,30,31). Two that are relevant for stratifying psychiatric disorders are 1) hierarchical clustering, which forms a hierarchy of cluster assignments by recursively splitting larger groups (“divisive clustering”) or combining individual samples (“agglomerative clustering” [e.g., Ward’s method (32)], and 2) community detection, which is a graph-based method that aims to cluster nodes into “communities” (33).

Finite Mixture Modeling

FMMs2 are a broad class of probabilistic approaches that aim to represent data using a finite number of parametric distributions (“components”). The simplest examples are Gaussian mixture models (GMMs),3 where all components have Gaussian distributions (24), but many other models are also members of this class (26), including latent class cluster analysis (LCCA) (25,34), growth mixture modeling (35), latent class growth analysis4 (LCGA) (36), and factor mixture modeling (20) (see Supplement).

LCCA is a particularly widely used approach that accommodates many different data types (e.g., continuous, categorical, and ordinal). It is highly generic and can model, for example, dependence between variables (e.g., to model correlated clinical variables) or can use covariates to help predict class membership (25,26,34). Growth mixture modeling is a useful generalization and is derived by combining FMM with growth models (26,35). This is appropriate for modeling longitudinal data derived from different growth trajectories. Given the neurodevelopmental basis for psychiatric disorders (37) and the importance of disease course in diagnosis (38), these approaches are increasingly being applied to stratify psychiatric disorders (39,40).

1We identified studies by performing a PubMed search for each disorder separately using the following search string: [clustering OR subtypes OR subgroups OR stratification] and (disorder name OR disorder acronym)]. We then selected a representative overview of studies for each disorder (this was exhaustive for ADHD, autism, and cross-diagnostic studies). For example, in the case of multiple studies using the same cohort, we only included the first or most important in this review. We also gave priority to studies that have not been reviewed previously (19,51,52).

2Many of the FMM approaches discussed here originate in the psychometric literature, which uses different nomenclature to mainstream statistics. Unfortunately, this nomenclature also varies between authors. We use consistent terminology throughout and synthesize with mainstream statistical literature wherever possible.

3Referred to as “latent profile analysis” in the psychometric literature.

4Also referred to as “group-based trajectory modeling.”
One advantage of FMMs is that they provide a full statistical model for the data, and therefore classical statistical techniques can be used to assess fit (e.g., likelihood ratio tests). They are also flexible: for example, GMMs can approximate any continuous distribution to acceptable error (41). However, modeling complex distributions may require many mixture components having many parameters.

Model Order Selection
Choosing the number of clusters or components is an important consideration and directly influences model flexibility. Many techniques have been proposed for comparing model orders, including classical information criteria (42,43) and specialized methods (44–46). Different methods embody different heuristics (e.g., how parameters are penalized), which may not yield the same or even a unique optimal model order, indicating that the data can be equally well-explained using different model orders. Some methods automatically estimate model order (33,49) but do not indicate whether other model orders are equally appropriate and often have additional parameters that influence the estimated model order. For example, graph-based methods (33) entail specifying a threshold above which nodes are considered connected (see Advantages and Disadvantages of Clustering for further discussion).

APPLICATIONS TO STRATIFY PSYCHIATRIC DISORDERS
Clustering methods have been used extensively to stratify all psychiatric disorders, both individually and across diagnoses; Tables 1–5 provide a representative (but not exhaustive) overview. Several articles offer more extensive quantitative reviews (19,50,51). Three salient observations can be made: first, during the many years that computational approaches have been used, relatively few algorithms have been used. There is, however, more variability among methods to select model order. Second, stratifications have been based on a range of measures, but predominantly symptoms or psychometric variables. This is notable considering that RDCC and ROAMER emphasize stratification on the basis of mappings between biological systems and cognitive domains, not just symptoms (19). To date, few studies have stratified psychiatric disorders on the basis of quantitative biological measures, and these studies have predominantly used neuroimaging-based measures (13,16,17,52). This may be because of well-known problems with clustering complex, high-dimensional data (see Advantages and Disadvantages of Clustering).

CLINICAL IMPLICATIONS
One of the most striking features evident from Tables 1–5 is that the outcomes of clustering are heavily dependent on the input data; the overall picture derived from the literature is a profusion of different ways to subtype psychiatric disorders with relatively little convergence onto a coherent and consistent set of subtypes (19,50). The disorder with the most consistent stratifications across studies is major depression, where many (53–56), but not all (57–59) studies report evidence for “typical” (melancholic) and “atypical” subtypes, although these often do not align with the classical DSM subtypes (60). In contrast, stratifications of schizophrenia, ADHD, and autism have been much more variable across studies. In these cases, it is difficult to know how these different clustering solutions relate to each other or which are most relevant for clinical decision-making. From a clinical perspective, the discrepancies in these findings may reflect different subgroupings being reflected in different measures or a convergence of multiple causal mechanisms on the same phenotype. There are hundreds of genetic polymorphisms associated with most psychiatric disorders (61,62), all having small effect sizes and converging on similar symptoms. This aggregation of small effects has been likened to a “watershed,” where genetic polymorphisms aggregate as they flow downstream, finding full expression in the syndromic expression of the disorder (63). An additional complication in comparing studies is that symptom profiles of many disorders vary over the course of the disorder, even within individual subjects (64). Therefore, quantitative comparisons between different studies and cohorts are needed, as is a greater focus on external validation (see below).

ADVANTAGES AND DISADVANTAGES OF CLUSTERING
Tables 1–5 show that clustering algorithms have been the method of choice for stratifying clinical groups and have made an important contribution to studying the heterogeneity underlying psychiatric disorders. Clustering methods are ideal if the disorder can be cleanly separated into subgroups (e.g., for separating typical from atypical depression). However, our review shows that psychiatric disorders cannot be reproducibly stratified using symptoms alone, probably because of extensive overlap between disorders. Indeed, finding an optimal solution is in general a computationally difficult problem (65). Therefore, all algorithms used in practice use heuristics to find approximate solutions that do not guarantee convergence to a global optimum. This is not overly problematic in itself, and standard approaches are to run multiple random restarts to find the best solution possible or to integrate different solutions to provide measures of cluster uncertainty. A more serious problem is that clustering algorithms always yield a result and partition the data into the specified number of clusters regardless of the underlying data distribution (Supplementary Figure S1). The number and validity of the clusters must be specified a priori or assessed post hoc. In this regard, it is important to recognize that different approaches to clustering embody different heuristics, possibly leading to different solutions. These heuristics are determined by many factors, including the choice of algorithm and distance function, the model order, the subspace in which clustering takes place, and the method used to search the

The overall objectives of clustering approaches and FMMs are similar; for the remainder of this article, we refer to both as “clustering” for brevity.

Technically, clustering belongs to the “NP-hard” class of problems.
<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects (N)</th>
<th>Measures</th>
<th>Algorithm</th>
<th>No. of Clusters (Method)</th>
<th>Cluster Descriptions</th>
<th>External Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmer et al., 1983</td>
<td>SCZ (65)</td>
<td>Symptoms and case history variables</td>
<td>K means and hierarchical clustering</td>
<td>2 (maximal agreement between methods)</td>
<td>Good premorbid adjustment, late onset, and well organized delusions. Poor premorbid functioning, early onset, incoherent speech, and bizarre behavior.</td>
<td>None</td>
</tr>
<tr>
<td>Castle et al., 1994</td>
<td>SCZ (447)</td>
<td>Symptoms and case history variables</td>
<td>LCCA</td>
<td>3 (χ² test)</td>
<td>Neurodevelopment. Paranoid Psychotic. Schizoaffective.</td>
<td></td>
</tr>
<tr>
<td>Murray et al., 2005</td>
<td>SCZ (387)</td>
<td>“Operational criteria” diagnostic measures (medical records and interview)</td>
<td>LCCA</td>
<td>BIC (42)</td>
<td>Depression. Reality distortion. Mania. Disorganization.</td>
<td>None</td>
</tr>
<tr>
<td>Cole et al., 2012</td>
<td>SCZ (208)</td>
<td>Social and academic adjustment scales</td>
<td>LCGA</td>
<td>3 [BIC and Lo-Mendell-Rubin test (42)]</td>
<td>Good—stable. Insidious onset. Poor deteriorating.</td>
<td>None</td>
</tr>
<tr>
<td>Bell et al., 2013</td>
<td>SCZ and SAD (77 + 63 validation)</td>
<td>Symptoms and social cognitive measures</td>
<td>K means</td>
<td>3 (Ward method)</td>
<td>High negative symptoms.                   High social cognition. Low social cognition.</td>
<td>None</td>
</tr>
</tbody>
</table>
space. Moreover, in general it is not possible to adjudicate unambiguously between methods because there is no clear measure of success for unsupervised learning methods. For example, different metrics for assessing model order often yield different answers and also may not identify a unique optimal model order. Therefore, heuristics and previous expectations play a strong role in the choice of algorithm and model order. Indeed, many studies use multiple approaches, aiming for consensus (Tables 1–5), but the final choice of method is often a matter of taste.

High-dimensional data bring additional problems for clustering that are well-recognized in the machine learning literature (see Supplementary Methods) (31,66). Specialized algorithms are therefore recommended for high-dimensional data (31,66), but to date these have not been applied to psychiatric disorders. Another problem for biological data (e.g., neuroimaging and genetics) is that the magnitude of nuisance variation is usually larger than clinically relevant variation, so the clustering solution can be driven by the nuisance variation rather than clinical heterogeneity. Therefore, it can be difficult to constrain clustering algorithms to find clinically relevant clusters, which necessitates careful data handling and preprocessing.

More specific problems with applying clustering algorithms to stratify psychiatric disorders include the following: 1) some participants may not clearly belong to any class; 2) some classes may be not well defined or may be unmanageably small (67); 3) subgroups may principally index severity (39,55,68); and 3) it is not clear whether healthy participants should be clustered separately or in combination with patients.

**VALIDATION**

The complexity of deriving clustering solutions makes validation crucial to ensure reproducibility and to ensure that the derived clusters index clinically meaningful variation. A common approach is to train supervised classifiers to separate classes using the same data that were used to derive the clusters or data that are highly correlated (e.g., different symptom measures). However, this approach is circular and simply measures how well classes can be separated within the training sample. A better approach is to assess cluster reproducibility, which requires additional cohorts or resampling of the data (e.g., cross-validation). However, to avoid bias, the entire procedure—including clustering—must be embedded within the resampling framework. To assess clinical validity, external data are necessary and should be defined a priori. For this, prediction of future outcome is considered the best test (69) if outcome can be clearly defined (e.g., the absence of relapse in schizophrenia). Biological measures can also provide useful validation because they can determine whether clusters map onto pathophysiology (11,12), which is important because subgroups that reduce phenotypic heterogeneity may not reduce biological heterogeneity (70).

7In contrast, there is a clear measure by which success of supervised methods can be assessed: the expected loss, measured by some loss function, over the joint distribution of labels and covariates. This can be estimated in various ways (e.g., cross-validation).
<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects (N)</th>
<th>Measures</th>
<th>Algorithm</th>
<th>No. of Clusters (Method)</th>
<th>Cluster Descriptions</th>
<th>External Validation</th>
</tr>
</thead>
</table>
| Paykel, 1971 (7) | Patients with depression (165) | Clinical interviews, case history, and personality variables | Friedman–Rubin algorithm (103) | 4 (maximize the ratio of between to within class scatter) | Psychotic
Anxious
Hostile
Young depressive with personality disorder | None |
| Maes et al., 1992 (57) | MDD (80) | Symptoms | K means | 2 (not specified) | Vital (i.e., psychomotor disorders, loss of energy, early morning awakening, and nonreactivity) Nonvital | Biological (e.g., endocrine) measures |
| Kendler et al., 1996 (63) | Female twin pairs (2163) | Symptoms | LCCA | 7 (not specified) | Only 3 clusters described:
Mild typical depression
Atypical depression
Severe typical depression | Body mass index, personality, and concordance of cluster membership among twin pairs |
| Sullivan et al., 1998 (54) | National comorbidity survey respondents (2836) | Symptoms | LCCA | 6 ($\chi^2$ statistic) | Severe typical
Mild typical
Severe atypical
Mild atypical
Intermediate
Minimal symptoms | Demographic and personality variables |
| Hybels et al., 2009 (58) | MDD (368) | Symptoms | LCCA | 4 [L$^2$ statistic (34), BIC] | DSM-IV depression: Moderate sadness, lassitude and inability to feel
Higher severity for all items, especially apparent sadness
Milder profile
Highest severity and most functional limitations | Demographic, social, and clinical variables |
| Lamers et al., 2010 (55) | MDD (818) | Symptoms plus demographic, psychosocial, and physical health variables | LCCA | 3 [BIC and AIC (43)] | Severe melancholic (decreased appetite, weight loss)
Severe atypical (overeating and weight gain)
Moderate severity | Stability over time, sociodemographic, clinical, and biological (e.g., metabolic) variables (104,105) |
| Lamers et al., 2012 (56) | National comorbidity survey—replication respondents. Adolescents (912) and adults (805) | Symptoms | LCCA | Adolescents: 3, adults: 4 (BIC) | Adolescents: Moderate typical
Severe typical
Severe atypical
Adults: Moderate
Moderate typical
Severe typical
Severe atypical | None |
Beyond Lumping and Splitting

Table 2. Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Measures</th>
<th>Algorithm</th>
<th>N</th>
<th>No. of Clusters (Method)</th>
<th>Cluster Descriptions</th>
<th>External Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhebergen et al., 2012 [74]</td>
<td>MDD (804) Longitudinal symptom scores</td>
<td>LCGA</td>
<td>5 (BIC and Lo-Mendell-Rubin test)</td>
<td>Remission, Decline (moderate severity), Decline (severe), Chronic (moderate severity), Chronic (severe)</td>
<td>External validation is defined as a data measure used to validate the derived classes that is of a different type to the data used to derive the classes. Wherever possible, we follow the own nomenclature for describing clusters.</td>
<td></td>
</tr>
<tr>
<td>Van Loo et al., 2014 [69]</td>
<td>MDD (8,261) Retrospective symptom scores and demographic data</td>
<td>K-means</td>
<td>3 (Inspection of dichotomization scores and area under the receiver operating characteristic curve [89])</td>
<td>High risk, Intermediate risk, Low risk</td>
<td>Functional magnetic resonance imaging scans, and clinical outcome.</td>
<td></td>
</tr>
<tr>
<td>Milaneschi et al., 2015 [70]</td>
<td>MDD (1477) Symptoms</td>
<td>LCCA</td>
<td>3 (BIC, AIC, and likelihood ratio test)</td>
<td>Severe melancholic, Moderate melancholic, Low risk</td>
<td>Functional magnetic resonance imaging scans, and clinical outcome.</td>
<td></td>
</tr>
</tbody>
</table>

Historically, the importance of validation has been somewhat overlooked (Tables 1–5), but it is reassuring to note that studies are increasingly validating stratifications against external measures, especially in the case of major depression (60,71–73); for example, Rhebergen et al. (39) derived a set of symptom trajectories to stratify depressed subjects that were subsequently validated against measures of affective processing derived from functional magnetic resonance imaging scans (73). Another notable example of external validation was provided by Karalunas et al. (14), who stratified children with ADHD on the basis of temperament ratings and validated these stratifications against cardiac measures, resting state functional magnetic resonance imaging scans, and clinical outcome.

ALTERNATIVES TO CLUSTERING

Surprisingly few alternatives to clustering have been proposed. Proposed alternatives are of 3 main types: first, some methods extend supervised learning to classify predefined disease states while accommodating uncertainty in the class labels. This has been achieved in the following ways: embedding the algorithm in a “wrapper” that identifies mislabeled samples [(74) Figure 1A, B]; semisupervised methods that only use labels for subjects with a definite diagnosis [(75) Figure 1C]; and hybrid methods that combine supervised learning with clustering [(76–78) Figure 1D] or fusing the image registration process with FMMs such that brain images are clustered at the same time as they are registered together (79). Second, manifold learning techniques (Figure 2A) have been used to find low-dimensional representations of the data that highlight salient axes of variation. For high-dimensional data, approaches that preserve local distances are well-suited for this [80] and have been used to find latent structure underlying neurologic disorders [81] and used for dimensionality reduction before clustering (82). Third, novelty detection algorithms, such as the one-class support vector machine [83], aim to identify samples that are different from a set of training examples [84] (Figure 3B).

Normative modeling (Figure 3) is an alternative approach for parsing heterogeneity in clinical conditions (22,85,86) and aims to model biological variation within clinical cohorts, such that symptoms in individual patients can be recognized as extreme values within this distribution. This can be compared to the use of growth charts to map child development in terms of height and weight as a function of age, where deviations from a normal growth trajectory manifest as outliers within the normative range at each age. This is operationalized by learning some decision function that quantifies the variation across the population range, including healthy functioning and also potentially symptoms (see Supplementary Methods). Such approaches have been proposed for identifying subjects that have an abnormal maturational trajectory in brain structure (86) or in cognitive development (85), or for mapping any clinically relevant variable (22). This approach breaks the symmetry inherent in case-control and clustering approaches and provides multiple benefits. First, it does not entail making strong assumptions about the clinical group (e.g., existence or number of subgroups). This was shown by Marquand et al. (22), where the clinical variables did not form clearly defined
Table 3. Studies Using Clustering Methods to Stratify Attention-Deficit/Hyperactivity Disorder

<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects (N)</th>
<th>Measures</th>
<th>Algorithm</th>
<th>No. of Clusters (Method)</th>
<th>Cluster Descriptions</th>
<th>External Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair et al., 2012 (15)</td>
<td>ADHD (285) and TDC (213)</td>
<td>Neuropsychologic scores</td>
<td>CD (33)</td>
<td>6 for ADHD (determined implicitly by the algorithm)</td>
<td>Response time variability (+)</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Working memory (-), memory span (-), inhibition (-), and output speed (-)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Working memory (-), memory span (-), inhibition (-), and output speed (-), minor differences in remaining measures</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Temporal processing (-)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Arousal (-)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Arousal (-), minor differences in remaining measures</td>
<td></td>
</tr>
<tr>
<td>Karalunas et al., 2014 (16)</td>
<td>ADHD (247) and TDC (190)</td>
<td>Personality measures (e.g., temperament)</td>
<td>CD</td>
<td>3 (determined implicitly by the algorithm)</td>
<td>Mild</td>
<td>Physiological (e.g., cardiac) measures, resting state fMRI and 1-year clinical outcomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Irritable (negative emotionality, anger, and poor soothability)</td>
<td></td>
</tr>
<tr>
<td>Gates et al., 2014 (16)</td>
<td>ADHD (32) and TDC (58)</td>
<td>fMRI (functional connectivity)</td>
<td>CD</td>
<td>5 (determined implicitly by the algorithm)</td>
<td>Subgroups characterized in terms of functional connectivity profiles</td>
<td>None</td>
</tr>
<tr>
<td>Costa Dias et al., 2015 (17)</td>
<td>ADHD (42) and TDC (63)</td>
<td>fMRI (reward related functional connectivity)</td>
<td>CD</td>
<td>3 (determined implicitly by the algorithm)</td>
<td>Subgroups characterized in terms of functional connectivity profiles</td>
<td>Clinical variables and reward sensitivity</td>
</tr>
<tr>
<td>Van Hulst et al., 2015 (67)</td>
<td>ADHD (96) and TDC (121)</td>
<td>Neuropsychological scores</td>
<td>LCCA</td>
<td>5 (BIC)</td>
<td>Quick and accurate</td>
<td>Parent ratings of behavioral problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Poor cognitive control</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Slow and variable timing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Remaining 2 groups were too small to characterize</td>
<td></td>
</tr>
<tr>
<td>Mostert et al., 2015 (106)</td>
<td>ADHD (133) and TDC (132)</td>
<td>Neuropsychological scores</td>
<td>CD</td>
<td>3 (determined implicitly by the algorithm)</td>
<td>Attention (-), inhibition (-)</td>
<td>Clinical symptoms and case history</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reward sensitivity (+)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Working memory (-) and verbal fluency (-)</td>
<td></td>
</tr>
</tbody>
</table>

External validation is defined as a data measure used to validate the derived classes that is of a different type to the data used to derive the classes. Wherever possible, we follow the authors’ own nomenclature for describing clusters, and a (+) or (–) indicates relative improvement or deficit in the specified variable.

ADHD, attention-deficit/hyperactivity disorder; BIC, Bayesian information criterion; CD, community detection; fMRI, functional magnetic resonance imaging; LCCA, latent class cluster analysis; TDC, typically developing control.
<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects (N)</th>
<th>Measures</th>
<th>Algorithm</th>
<th>No. of Clusters (Method)</th>
<th>Cluster Descriptions</th>
<th>External Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Munson et al., 2008 (107)</td>
<td>ASD (245)</td>
<td>IQ scores</td>
<td>LCCA and taxonometric analysis</td>
<td>4 (BIC, entropy, and Lo-Mendell-Rubin test)</td>
<td>Low IQ, Low verbal IQ/medium nonverbal, Medium IQ, High IQ</td>
<td>Symptom scores</td>
</tr>
<tr>
<td>Sacco et al., 2012 (21)</td>
<td>ASD (245)</td>
<td>Demographic, clinical, case history, and physiologic (e.g., head circumference) variables</td>
<td>K means</td>
<td>4 (Ward’s method)</td>
<td>Immune + circadian and sensory, Circadian and sensory, Stereotypic behaviors, Mixed</td>
<td>None</td>
</tr>
<tr>
<td>Fountain et al., 2012 (40)</td>
<td>ASD (6795)</td>
<td>Symptoms</td>
<td>LCGA</td>
<td>6 (BIC)</td>
<td>High functioning, Demographic variables and autism risk factors</td>
<td></td>
</tr>
<tr>
<td>Georgiades et al., 2013 (108)</td>
<td>ASD (391)</td>
<td>Symptom scores</td>
<td>FMM</td>
<td>3 (AIC and BIC)</td>
<td>Social communication (–), repetitive behaviors (+)</td>
<td>Demographic variables and cognitive measures</td>
</tr>
<tr>
<td>Doshi-Velez et al., 2014 (109)</td>
<td>ASD (4927)</td>
<td>Electronic medical records</td>
<td>Ward’s method</td>
<td>4 (Ward’s method)</td>
<td>Seizures, Multisystem disorders, Auditory disorders and infections, Psychiatric disorders, Not otherwise specified</td>
<td>None</td>
</tr>
<tr>
<td>Veatch et al., 2014 (68)</td>
<td>ASD (1261 + 2563 for replication)</td>
<td>Symptoms, demographic, and somatic variables</td>
<td>Ward’s method</td>
<td>2 [Adjusted Arabie Rand index (110) and validation with additional clustering algorithms]</td>
<td>Severe, Less severe</td>
<td>Genomic data</td>
</tr>
</tbody>
</table>

External validation is defined as a data measure used to validate the derived classes that is of a different type to the data use to derive the classes. Wherever possible, we follow the authors’ own nomenclature for describing clusters, and a (+) or (–) indicates relative improvement or deficit in the specified variable. 

ASD, autism spectrum disorder; BIC, Bayesian information criterion; FMM, factor mixture modeling; LCCA, latent class cluster analysis; LCGA, latent class growth analysis.
clusters but normative modeling identified distinct brain mechanisms that give rise to symptoms. Second, it allows both normal functioning and deviations from normal functioning that may underlie symptoms to be mapped in individual subjects. Third, it permits diagnostic labels to be used as predictor variables, enabling inferences over the labels. Finally, it intuitively matches the clinical conception where diseases in individual patients are recognized as deviations from normal functioning. This approach can be used to estimate mappings between biology and behavior across multiple cognitive domains; therefore, it is well aligned with RDoC and ROAMER and also complements clustering because clustering algorithms can still be applied to these mappings. On the other hand, normative modeling requires careful data processing to ensure that the outliers detected are not outliers from the normative distribution due to artifacts. It is also best suited to large normative cohorts that capture the full range of functioning in the reference population.

**DISCUSSION**

In this article, we introduced the basic concepts of data-driven stratification of psychiatric disorders and reviewed the existing literature. The overwhelming majority of studies have employed clustering or FMM, aiming to subgroup clinical populations. This has been somewhat successful (Tables 1–5), although the results are heavily dependent on the type of data used; for most disorders, both the number and characteristics of the derived clusters vary between studies, and a consensus as to a consistent set of subgroups is yet to be reached. We highlighted the importance of validation to ensure that derived clusters map onto clinically relevant variation and outlined various alternatives to clustering.

The ongoing discussion surrounding psychiatric nosology reflects well-acknowledged difficulties in finding biological markers that predict current disease state or future outcomes with sufficient sensitivity and specificity to be clinically useful (1,10). While this is an important motivation behind RDoC and ROAMER (11,12,87), this review highlights that neither the reclassification of psychiatric disorders nor the emphasis on cutting across current diagnostic classifications is a central innovative feature. A more important contribution is a shift away from symptoms and towards conceptualizing pathology as spanning multiple domains of functioning and across multiple levels of analysis. In RDoC, this is represented as a matrix with rows containing basic cognitive dimensions (“constructs”) grouped into domains of functioning (e.g., positive or negative valence systems) and columns containing units of analysis (e.g., genes, cells, or circuits) (87). Viewed in this light, clustering of algorithms provides only a partial answer to the challenges posed by RDoC and ROAMER because it does not provide an obvious means to link constructs with units of analysis. Put simply, it is necessary to link the rows of the RDoC matrix with its columns and chart the variation in these mappings. This is necessary before the clinical validity of RDoC domains can be assessed as to whether they predict disease states more accurately than classical diagnostic categories (38).

Surprisingly few methods have been proposed that meet these objectives. Most that do exist aim to break the symmetry that both the case-control paradigm and clustering approaches entail in that all clinical groups are well-defined entities. Normative modeling (22,85,86) is one particularly promising approach that aims to map variation in clinically relevant variables, so that each individual subject can be placed within the population range and disease can be considered as an extreme deviation from a normal pattern of

---

**Table 5. Studies Employing Clustering Methods to Stratify Patients in a Cross-Diagnostic Setting**

<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects (N)</th>
<th>Measures</th>
<th>Algorithm</th>
<th>No. of Clusters (Method)</th>
<th>Cluster Descriptions</th>
<th>External Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olino et al., 2010 (113)</td>
<td>Adolescents (1653), including MDD (603), ANX (253), SUD (453)</td>
<td>Diagnosis (longitudinal)</td>
<td>LCGA</td>
<td>6 (BIC)</td>
<td>Persistent depression, Persistent anxiety</td>
<td>Demographic and case history variables</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Late onset anxiety, increasing depression</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Increasing depression</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Initially high, decreasing anxiety</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Absence of psychopathology</td>
<td></td>
</tr>
<tr>
<td>Lewdanowski et al., 2014 (111)</td>
<td>SCZ (41), SAD (53), BPDp (73)</td>
<td>Clinical and cognitive measures</td>
<td>K means</td>
<td>4 (Ward’s method)</td>
<td>Neuropsychologically normal, Diagnosis, demographic variables, and community functioning</td>
<td>Mixed cognitive profiles (X:2)</td>
</tr>
<tr>
<td>Kleinman et al., 2015 (112)</td>
<td>ADHD (23), BPD (10), BPDa (53), and HCs (18)</td>
<td>Continuous performance test measures</td>
<td>K means</td>
<td>2 [Silhouette index (46)]</td>
<td>Sustained attention (+), inhibitory control (+), impulsiveness (+), and vigilance (-)</td>
<td>Diagnosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The converse of above</td>
<td></td>
</tr>
</tbody>
</table>
functioning. This provides a workable alternative to lumping and splitting the psychiatric phenotype and a method to chart variability across different domains of functioning and different units of analysis.

Our review also highlighted that few studies have used biological measures to derive stratifications. This may be because of difficulties that unsupervised methods have with separating nuisance variation from clinically relevant variation, particularly in high dimensions (31). This may be particularly problematic in genomic studies; some reports have used genomic data as validation of the derived clusters (60,68), but the only study we are aware of that used genomic data to derive clusters (88) has received severe criticism for inadequately dealing with artefactual variation.8

One way that this problem may be addressed in the future is by developing richer clustering models that integrate clinical or domain knowledge in a way that guides the clustering algorithm toward clinically relevant variation. A simple example is the use of growth mixture models to cluster samples on the basis of within-participant change over time (39,40). More generally, probabilistic graphic models (24) provide an elegant framework that allows existing knowledge to be incorporated to help find clinically meaningful clusters. To our knowledge, this approach has not been used in psychiatry, but it has been useful to stratify disease cohorts in other clinical domains (89). Other emerging machine learning techniques that may be fruitfully applied to stratifying psychiatric disorders include probabilistic methods that allow for multiple labels within individual patients (90), clustering methods that do not uniquely assign points to a single cluster (31), and deep learning methods (91,92).

8For example, see the discussion at: http://www.ncbi.nlm.nih.gov/pubmed/25219520.

Figure 1. Schematic examples of alternative approaches to clustering and finite mixture models based on supervised learning. (A) This example shows the benefit of correcting mislabeled training samples. A supervised classifier trained to separate experimental classes (black and red points) may be forced to use a complex nonlinear decision boundary (blue line) to separate classes if data points are mislabeled (circled). (B) A simpler decision boundary results if the incorrect labels are corrected, for example using a wrapper method (74). (C) In a semisupervised learning context (75), only some data points have labels (black and red points). These can correspond to samples for which a certain diagnosis can be obtained. All other data points are unlabeled, but can still contribute to defining the decision boundary. Hybrid methods (76–78) combine supervised classification with unsupervised clustering and use multiple linear decision boundaries to separate the healthy class (blue points) from putative disease subgroups (colored points). See text for further details.
In summary, we reviewed the literature for stratifying psychiatric disorders and showed that the field has, to date, relied heavily on clustering and FMM. These undoubtedly provide an important contribution but only partially satisfy the objectives of RDoC and ROAMER. It is also necessary to chart variation in brain-behavior mappings to fully parse heterogeneity across domains of functioning and diagnostic categories. The hope is that using such mappings to derive future disease stratifications will enable clinical phenotypes to be dissected along the most relevant axes of variation, ultimately enabling treatments to be better targeted to individual patients.

ACKNOWLEDGMENTS AND DISCLOSURES
This work was supported by the Netherlands Organization for Scientific Research (NWO) under the Gravitation Programme (Grant No. 024.001.006 supporting AFM), by a Marie Curie International Incoming Fellowship under the European Union’s Seventh Framework Programme (Grant No. FP7/2007-2013) Grant No. 227340 (BRAIN FINGERPRINT to MM), and by a VIDI grant from the NWO (Grant No. 864-12-003 to CFB). JB received funding from the FP7 under Grant Nos. 602805 (AGGRESSOTYPE), 603016 (MATRICS), and 278948 (TACTICS) and from the European Community’s
Horizon 2020 Programme (H2020/2014-2020) under Grant Nos. 643051 (MiND) and 642996 (BRAINVIEW). We also acknowledge funding from the Wellcome Trust UK Strategic Award (098369/Z/12/Z).

JB has been a consultant to, advisory board member of, and a speaker for Janssen Cilag BV, Eli Lilly, Shire, Lundbeck, Roche, and Servier. He is not an employee of any of these companies, and not a stock shareholder of any of these companies. He has no other financial or material support, including expert testimony, patents or royalties. CFB is director and shareholder of SBGneuro Ltd. The other authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION
From the Donders Centre for Cognitive Neuroimaging (AFM, TW, MM, JB, CFB), Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen; Department of Cognitive Neuroscience (AFM, JC, CFB), Radboud University Medical Centre, Nijmegen; and Karakter Child and Adolescent Psychiatric University Centre (JB), Nijmegen, The Netherlands; Department of Neuroimaging (AFM), Centre for Neuroimaging Sciences, Institute of Psychiatry, King’s College London, London; and Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (CFMRIB), University of Oxford, London, United Kingdom.

Address correspondence to Andre F. Marquand, Ph.D., Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Kapittelweg 29, 6525 EN, Nijmegen, The Netherlands; E-mail: a.f.marquand@donders.nl.

Received Jan 28, 2016; revised Apr 6, 2016; accepted Apr 6, 2016.

Supplementary material cited in this article is available online at http://dx.doi.org/10.1016/bpsc.2016.04.002.

REFERENCES

Beyond Lumpining and Splitting


