



# Phenomapping: Methods and Measures for Deconstructing Diagnosis in Psychiatry

# 7

Andre F. Marquand, Thomas Wolfers, and Richard Dinga

In most areas of medicine, the advent of biological tests to measure disease state has revolutionised diagnosis and treatment allocation. However, this is not the case in psychiatry, which is now virtually the last area of medicine where diseases are still diagnosed based on symptoms and biological tests to assist treatment allocation remain to be developed (Kapur et al. 2012). This is especially problematic because psychiatric disorders are all extremely heterogeneous, both in terms of their clinical presentation (which we refer to as ‘clinical heterogeneity’), in terms of their underlying biological causes (‘biological heterogeneity’) and in terms of environmental factors (‘environmental heterogeneity’). Even though diagnostic criteria have been periodically revised over the years, these sources of heterogeneity remain a substantial barrier to better understanding the causative mechanisms of psychiatric disorders and to developing optimal treatments. Indeed, there have been virtually no new therapeutic targets in psychiatry for decades.

---

A. F. Marquand

Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Department of Cognitive Neuroscience, Radboud University Medical Centre, Nijmegen, The Netherlands

Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, King’s College London, London, UK

e-mail: [a.f.marquand@fcdonders.ru.nl](mailto:a.f.marquand@fcdonders.ru.nl)

T. Wolfers

Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Department of Human Genetics, Radboud University Medical Centre, Nijmegen, The Netherlands

R. Dinga

Department of Psychiatry, Amsterdam Neuroscience and Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands

The overwhelmingly dominant paradigm in psychiatric research has been the case-control approach, which assumes that patient and control groups each form a distinct entity and completely ignores heterogeneity within cohorts. It has long been recognized that we must look beyond simple case-control comparisons to be able to deconstruct the heterogeneous phenotype of psychiatric disorders and, correspondingly, there have been many attempts to find data-driven reclassifications or stratifications of psychiatric disorders (Marquand et al. 2016b; Schnack 2018). The dominant approach has been to train unsupervised machine learning algorithms on the basis of symptoms or psychometric variables aiming to find data-driven subtypes of patients. Like the case-control approach, this assumes that patient cohorts can be cleanly partitioned into distinct subtypes. However, despite more than 40 years of effort, this approach has not converged upon a reproducible and clinically useful set of subtypes for any psychiatric disorder (Marquand et al. 2016b). Frustration with this lack of progress has led to several large-scale initiatives that aim to use many different biological and behavioural measures to finally bring the era of ‘precision medicine’ to psychiatry (Insel and Cuthbert 2015). The most prominent of these are the Research Domain Criteria (RDoC) initiative proposed by the National Institute of Mental Health in the USA (Insel et al. 2010) and the European Roadmap for Mental Health Research (ROAMER) (Schumann et al. 2014). The central feature of these initiatives is to move away from using only symptoms for disease classifications and instead to integrate biological and behavioural measures from different levels of analysis (e.g. genes, cells and circuits) and across different domains of functioning (e.g. positive affect, social processing). Although the short-term focus of RDoC and ROAMER is principally on research, the clear implication is that the current nosological classifications will eventually need to be revised. The way this is most popularly envisaged to occur is that by integrating across domains of functioning and across different biological and behavioural levels, psychiatric cohorts will be cleanly separable into subtypes that simultaneously cut across current diagnostic classifications and relate more closely to underlying brain systems (Insel and Cuthbert 2015). At the time of writing, it is eight years since RDoC was officially released, and it is fair to say that RDoC and similar initiatives have, to date, also provided only a modest yield. Whilst RDoC has driven substantial basic research,<sup>1</sup> there are still few successful attempts to stratify psychiatric disorders on the basis of biological systems and none that are close to challenging the current diagnostic criteria in clinical practice.

In view of the considerations above, in this chapter we will review the literature aiming to employ biological measures to stratify the phenotype of psychiatric disorders. First, we will briefly review the biological measures that useful for stratifying patient cohorts. Second, we give a brief conceptual overview of the different methodological approaches that have been employed for this purpose. Third, we will provide a focused review of studies that have used biological measures to derive stratification, in line with RDoC and ROAMER. Finally, we will

---

<sup>1</sup>See e.g. <https://www.nimh.nih.gov/research-priorities/rdoc/nimh-rdoc-publications.shtml>.

identify difficulties to finding reproducible and clinically meaningful stratifications and suggest new directions for the field. We will argue that a fixation on simple case-control type differences between well-defined subgroups has been a major limiting factor in finding reproducible and clinically meaningful stratifications.

---

## 7.1 Measuring Biology in Big Data Cohorts

In recent years clinical neuroscience has undergone a tectonic shift away from small, boutique studies towards big data cohorts. This entails an enormous increase both in the number of different measures of biology and behavior that are acquired and also in the size of the cohorts from which they are derived. For example, in genetics, large international consortia and data sharing initiatives have emerged that are providing increasing numbers of genome-wide significant hits for psychiatric disorders (e.g. Ripke et al. 2014). However, the effect size of all individual genetic variants discovered to date are small and even aggregation of many effects through polygenic risk scores only explains a tiny fraction of the variance in the phenotype of psychiatric disorders (e.g. Milaneschi et al. 2015). This means that genetic measures are probably better suited to profiling and validating prospective stratifications rather than deriving the stratifications themselves. At the same time, advances in brain imaging techniques now make it possible to measure many aspects of brain structure, function and connectivity non-invasively and in vivo. There are also now many large population-based studies that acquire a range of neuroimaging, behavioural and clinical measurements from large cohorts (e.g. the UK Biobank study (Miller et al. 2016) and the Human Connectome Project (Van Essen et al. 2013)). Together, this makes neuroimaging the most widely used—and arguably most promising—method for deriving biologically based stratifications of psychiatric disorders. However, many other measures also provide promising and potentially complimentary information for this purpose; for example, blood-based biomarkers (Lamers et al. 2013), continuous behavior monitoring from smartphones and wearable sensors (Torous et al. 2017) or electronic monitoring of continuous speech patterns (Bedi et al. 2015) but at the present time, these remain relatively unexplored for the purposes of stratification. Of course, different measures can also be combined via multi-modal data fusion (e.g. Wolfers et al. 2017), at the expense of increasing the complexity of the analytical pipeline. Consequently, the time has never been better for the application of machine learning based methods for data-driven stratification of psychiatric disorders on the basis of biological readouts. However, the advent of big data for clinical neuroscience brings particular analytical challenges. These include difficulties in scaling off-the-shelf approaches to high dimensional problems (Kriegel et al. 2009) and developing methods to capture clinically relevant variation across large cohorts of participants whilst separating that variation from nuisance variation (e.g. due to artefacts or site effects). Meaningful stratification of psychiatric disorders is therefore heavily dependent on the underlying analytical methodology.

## 7.2 Overview of Analytical Approaches for Stratification

The overwhelming majority of applications of machine learning methods to big data psychiatry have been *supervised* in the sense that they are provided with labels and the learning process consists of estimating a mapping between inputs (e.g. biomarkers) and outputs (e.g. diagnostic labels). There are many different approaches for supervised learning, including support vector machines (Boser et al. 1992), penalized linear models (Hastie et al. 2009) Bayesian approaches (Rasmussen and Williams 2006) and deep learning (LeCun et al. 2015). Whilst these differ with regard to the underlying model assumptions, associated estimation procedures and the accuracy with which they can predict the target labels, the fundamental idea behind all these approaches is the same in that the algorithm seeks to maximize the accuracy of predicting the label of new data points (Fig. 7.1a). In psychiatry, supervised learning has been widely used both for predicting diagnosis (Wolfers et al. 2015) and quantitative psychometric variables (e.g. Mwangi et al. 2012) on the basis of neuroimaging biomarkers.

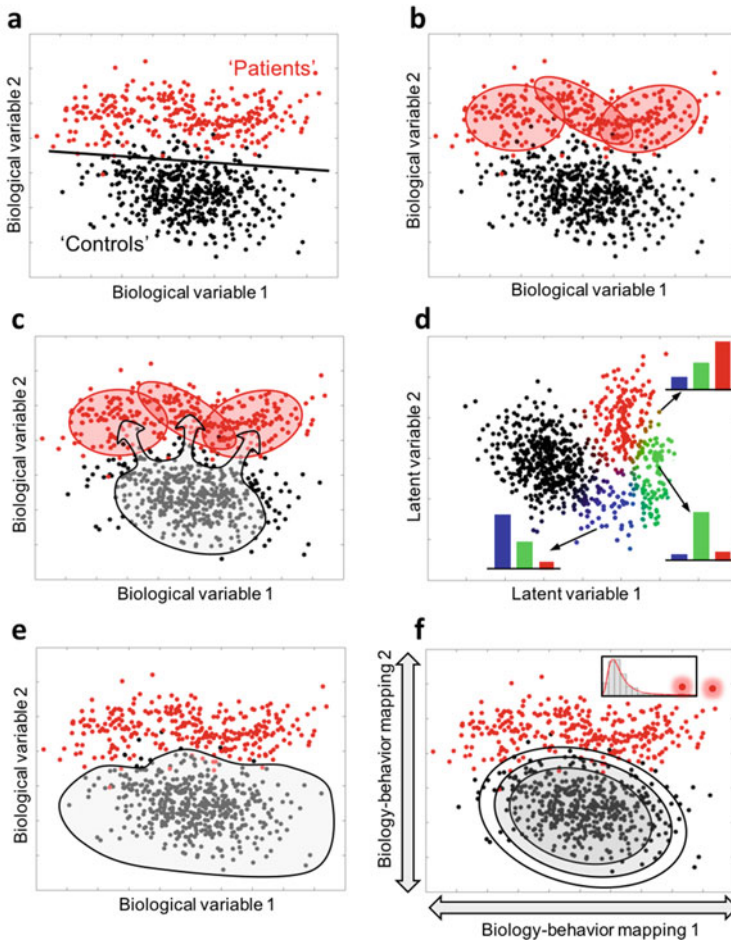
The supervised approach is reasonable if the labels are known in advance and are both accurate and reliable. However, in psychiatry labelling errors are probably relatively common (e.g. due to clinical or biological heterogeneity in addition to misdiagnosis or comorbidity). With this in mind and since the aim of stratification is to understand variation within the disease group (i.e. independently from the diagnostic labels), supervised learning is not widely used for stratifying disease groups. One exception is supervised learning methods that include mechanisms for correcting errors in the labels (e.g. Young et al. 2013), which may be useful to identify atypical samples.

In contrast, in *unsupervised* learning, the machine learning algorithm is not provided with target values and learns to find structure in the data by applying heuristics encoded in each algorithm to the data. There are many types of unsupervised learning algorithm, including clustering, matrix factorization methods, latent variable models and anomaly detection methods (Hastie et al. 2009). Unsupervised learning approaches are often suitable for exploratory data analysis and are, on the face of it, well suited to stratifying the phenotype of psychiatric disorders and are widely used for this purpose (Marquand et al. 2016b; Schnack 2018).

---

## 7.3 Clustering

Clustering algorithms are probably the most widely used unsupervised approach in general and are certainly the most widely used methods for stratifying psychiatric disorders. The central idea is that an algorithm is trained to partition a set of data points (i.e. subjects) into different clusters on the basis of some measurements (e.g. derived from neuroimaging data), such that the samples in each cluster are more similar in some sense to one another than to those in the other clusters (Fig. 7.1b). This entails defining a measure of similarity or distance between data points (e.g.



**Fig. 7.1** Schematic overview of different approaches to parsing heterogeneity in psychiatric disorders on the basis of biological data. **(a)** Supervised learning approaches regard the patient and control groups as distinct entities, and thereby ignore heterogeneity within the data. **(b)** Clustering algorithms aim to partition one or both of the groups into discrete clusters. Here a Gaussian mixture model was estimated to partition the patient group into three clusters. Shown are the ellipsoids corresponding to one standard deviation from the cluster centers. **(c)** A hybrid method that combines clustering and distribution matching (Dong et al. 2016). Here the method estimates a set of three transformations that match the distribution of the control group to the distribution of the patient group. **(d)** A latent variable model that models symptoms as arising from a set of three latent disease processes (e.g. Zhang et al. 2016). The data are represented according to a set of latent variables (of which only two are shown as axes). Each datapoint from the patient group is colored according to the proportion of each latent process it expresses via red, blue or green hue. The loadings for three hypothetical data points are shown. **(e)** Outlier detection method that estimates a decision boundary enclosing the control group, aiming to detect patients as outliers (Mourao-Miranda et al. 2011). **(f)** Normative modelling approaches aim to estimate a normative distribution over a reference cohort such that the abnormality of each individual participant can be quantified via extreme value statistics. The extreme value abnormality score for one datapoint is shown in the inlay along with a fit extreme value distribution. Note that the normative distribution is defined with respect to a set of mappings between biology and behavior, analogous to ‘growth charts’ in somatic medicine (Marquand et al. 2016a). See text for further details

Euclidean distance or correlation) and the desired number of clusters. In the present work, we largely gloss over the differences between different clustering algorithms (e.g. K-means clustering, finite mixture modelling and graph-based clustering) and label these approaches simply as ‘clustering’. We refer the interested reader to our previous work for more detail, where we provide a detailed introduction to some common clustering algorithms along with methodological considerations relating to their implementation (Marquand et al. 2016b).

---

## 7.4 Studies Subtyping Psychiatric Disorders on the Basis of Biology (‘Biotyping’)

As noted above, most applications to stratify psychiatric disorders on the basis of biology are based on the application of off-the-shelf clustering techniques, where the derived clusters are sometimes referred to as ‘biotypes’ (e.g. Clementz et al. 2016; Drysdale et al. 2017). One thing immediately apparent from a survey of the literature is a paucity of studies that report stratifications derived from biological measures, especially relative to the proliferation of applications of clustering algorithms to psychometric data (Marquand et al. 2016b). This is perhaps surprising given the strong motivation provided by the tight integration of research funding with initiatives such as RDoC and ROAMER (Insel et al. 2010; Schumann et al. 2014). One reason for this may be that biological data are often complex and high-dimensional with many different axes of variance. Clustering is a notoriously difficult problem in high dimensions (Kriegel et al. 2009) because many axes of variance may be artefactual or irrelevant and different axes may be important for different clusters within the same clustering solution. As a result, most applications reviewed here employ extreme dimensionality reduction, often training clustering algorithms on as few as two dimensions or alternatively use parameters from other models as features for clustering.

One of the earliest efforts to derive biotypes for stratifying psychiatric disorders was provided by Brodersen et al. (2014) who stratified a cohort of schizophrenia patients using Bayesian mixture model on the basis of parameters derived from a model of working memory estimated from functional magnetic resonance imaging (fMRI) data. This yielded three patient subgroups which differed in terms of symptom severity. Another study used structural connectivity measures derived from diffusion tensor imaging to stratify patients with first episode schizophrenia (Sun et al. 2015). This study reported two subtypes, which differed in terms of their profile of white matter abnormalities and symptom profile.

In a prominent study by Clementz et al. (2016), the authors derived a set of three biotypes from large cohort of patients with psychosis spectrum disorders using a broad panel of biomarkers, including neuropsychological, saccadic control and electroencephalography measures. These subtypes cut across classical diagnostic boundaries and had distinctive patterns of grey-matter reductions in a graded fashion such that one of the biotypes had patterns of reduction intermediate between the other two, a pattern also evident in relatives of the probands. Brain structural

differences were further explored in a follow-up study (Ivleva et al. 2017), but since these analyses were performed on the same cohort, this cannot be considered a replication.

Another prominent study reported finding four biotypes of depression on the basis of mappings between resting state fMRI connectivity measures and symptoms derived from a multi-site cohort (Drysdale et al. 2017). These biotypes again crossed classical diagnostic boundaries and had differential characteristics with regard to symptoms and fMRI connectivity. The authors of this study performed limited validation of these subtypes on additional data samples and also demonstrated that the derived subtypes predicted treatment response (trans-cranial magnetic stimulation).

Finally, two studies from the same group have aimed to stratify attention-deficit hyperactivity disorder (ADHD) using functional connectivity measures derived from on fMRI (Gates et al. 2014; Costa Dias et al. 2015). These reported different numbers of clusters (3 and 5, respectively), and characterized the different subtypes in terms of their connectivity profiles although in the case of (Costa Dias et al. 2015), these were also related to symptom severity. As noted by the authors of these studies, this highlights that there are always multiple ways to partition cohorts using clustering algorithms, even based on the same data. These alternative solutions may be equally valid, for example when assessed according to different metrics (see below for further discussion).

---

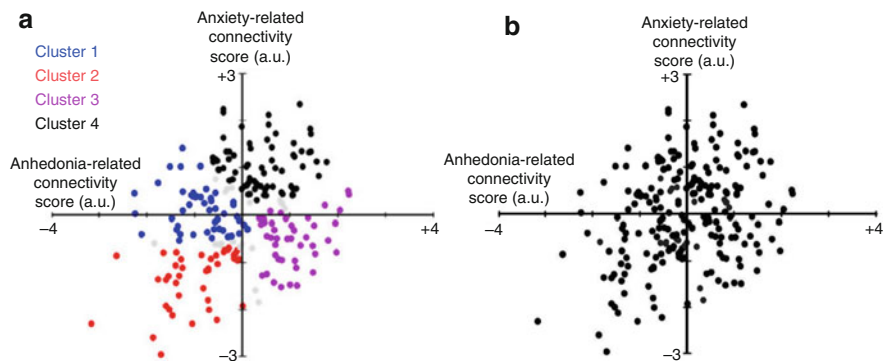
## 7.5 Alternatives to Biotyping

There are multiple alternative analytical approaches for stratifying psychiatric disorders including hybrid methods that combine supervised learning with clustering (Varol et al. 2017), hybrid methods that combine distribution matching and clustering (Dong et al. 2016), methods that model the emergence of symptoms in individual subjects as deriving from a linear combination of latent disease factors (Ruiz et al. 2014; Zhang et al. 2016), outlier or anomaly detection methods (Mourao-Miranda et al. 2011) and normative modelling techniques that aim to chart variation in population cohorts and place each individual subject within the population range (Marquand et al. 2016a).

For example, the method proposed in (Dong et al. 2016) is a hybrid of clustering and distribution matching. This method was explicitly designed for structural brain imaging data and tackles heterogeneity within the patient cohort by training an algorithm that estimates a discrete set of transformations that warp the distribution of control participants to match the patient distribution (Fig. 7.1c). The intuition is that each of the different transformations encodes a different biotype. The method also provides a posterior probability measure quantifying the certainty with which each datapoint belongs to each biotype or, in other words, it provides a ‘soft’ clustering of the data. This was used to stratify a cohort of schizophrenia patients on the basis of structural MRI data (Honnorat et al. 2018), yielding three subtypes with different patterns of volumetric difference relative to control subjects.

Another alternative approach is based on the assumption that each individual expresses a set of latent disease factors to varying degrees, which together comprise an individualized symptom profile (Ruiz et al. 2014; Zhang et al. 2016). Such methods can be seen as relaxing the requirement that each subject belongs to a single cluster or subtype (Fig. 7.1d). A particularly promising approach along this line is topic modelling, which describes a collection of natural language processing techniques that aim to find a set of topics that occur frequently in a collection of documents such that each document is assumed to relate to multiple topics. For example, in (Zhang et al. 2016) the authors applied a common topic modelling technique—latent Dirichlet allocation (LDA; Blei et al. 2003)—to stratify Alzheimer’s disease patients on the basis of structural MRI. In contrast to clustering approaches, LDA models disease in each individual patient (analogous to a ‘document’) as emerging from a pre-specified number of latent disease processes (‘topics’), which are expressed to different degrees in different patients. Typically, LDA is framed as a probabilistic model, which can readily yield quantities of interest such as the probability that a given individual expresses a particular latent disease factor. In, the study by Zhang and colleagues (Zhang et al. 2016), the authors discovered three hierarchical latent disease factors characterized by different patterns of atrophy and different trajectories of cognitive decline.

In contrast, anomaly or outlier detection methods aim to estimate a predictive function or decision boundary that characterizes the support of the distribution of a healthy class. The intuition then is that ‘abnormal’ samples can be detected as outliers (Fig. 7.2e). Probably the most common approach in neuroimaging is the one-class support vector machine (OC-SVM; Sato et al. 2012). For example,



**Fig. 7.2** Clustering algorithms can impose artificial categorical structure on underlying continuous variation. (a) Clustering solution from a study stratifying depression on the basis of symptoms and brain functional connectivity data (reproduced with permission from Drysdale et al. 2017). Each axis describes subject level loadings from canonical correlation analysis. Different colors represent different clusters and gray clusters are ambiguous data points that were excluded from the analysis when computing the distinctiveness of each cluster. (b) The same data with the cluster labels removed. It is clear that the evidence for clusters in the data is equivocal. The data could be equally well—and probably better—explained using a continuous model along two dimensions



Mourao-Miranda et al. (2011) applied this approach to fMRI data derived acquired while healthy participants and patients with depression were performing an affective processing task. The algorithm detected patients as outliers such that the degree of abnormality detected correlated with depression symptoms. The OC-SVM can be applied to relatively high dimensional problems, but only provides a decision boundary. In other words, it does not characterize the distribution statistically, nor provide estimates of variation within the distribution. For this, density estimation techniques (Hastie et al. 2009) could theoretically be applied, but these are largely limited to low dimensional problems.

Normative modelling (Marquand et al. 2016a; Fig. 7.2f) is an emerging statistical technique that approaches the stratification problem from a different perspective. Under this framework, a statistical model is estimated to chart centiles of variation in clinical cohorts such that each individual patient can be placed within the population distribution. This is analogous to the use of growth charts in somatic medicine to map child development (e.g. in terms of height or weight) as a function of age. At the heart of normative modelling is the estimation of mappings between psychometric variables and a quantitative biological readout that provide estimates of variation across the population. A straightforward example of such a mapping would be between chronological age and brain structure to form a 'brain growth chart', which is useful because most psychiatric disorders are rooted in an underlying trajectory of brain development (Insel 2014). However, the method is agnostic to the type of measures that are employed and it can be used to chart variation along any biological-behavioural axis. For example, in (Marquand et al. 2016a) a normative model of reward processing was estimated linking behavioural measures of delay discounting with reward-related brain activity. A second key ingredient in normative modelling is the use of extreme value statistics (Beirlant et al. 2004) to perform statistical inference over the aspects of the pattern that are most abnormal. The intuition behind this is that the method focusses on the most extreme differences from the expected pattern, following the notion that those differences are those most likely to be implicated in disease. In contrast, most of the more prevalent statistical techniques (e.g. t-statistics and analyses of variance) focus on central tendency, which is useful to detect mean differences between groups of participants but has limited ability to provide inferences about either individual participants, or about the aspects of the pattern that are most abnormal. The third key ingredient in normative modelling is the choice of the reference cohort. The most straightforward choice is to select only healthy participants such that deviations from the normative model can be interpreted as deviations from a healthy pattern. However, a different reference cohort could also be chosen, which includes subjects with different diagnoses as well as healthy subjects. If the prevalence of the different disorders within such a cohort matches the population prevalence, then such a cohort provides an accurate reflection of how abnormalities can be interpreted with respect to the population at large, which is often of interest in an epidemiological context.

Normative modelling has several distinguishing characteristics that set it apart from other methods. First, it provides statistical measures of deviation from a healthy pattern for each individual subject, in other words, providing personalized

statistical predictions or ‘fingerprints’ that are at the heart of precision medicine (Insel and Cuthbert 2015; Kapur et al. 2012; Mirnezami et al. 2012). Second, normative modelling is completely agnostic to the diagnostic labels, which means they can be included as predictor variables to explain variance in the reference cohort. This is important because we must not overlook the discriminative power of diagnosis in many cases (Weinberger and Goldberg 2014). Third, normative modelling does not require that subjects share similar or overlapping patterns of abnormality and does not assume that the clinical cohort can be cleanly partitioned into subgroups although clustering algorithms can of course be trained on the deviations derived from normative models. This means it is useful to understand the variance structure in clinical cohorts where there are no clearly defined subtypes (e.g. where pathology may be better described as following a spectrum of functioning). In line with these considerations, some early application of normative modelling in schizophrenia, bipolar disorder, attention-deficit/hyperactivity disorder and autism spectrum disorders on the basis of structural MRI (Wolfers et al. 2018 <https://www.ncbi.nlm.nih.gov/pubmed/30304337>, <https://www.biorxiv.org/content/early/2018/11/27/477596>) are showing that group-level difference—or in other words differences in the ‘average patient’—are only the ‘tip of the iceberg’. Instead, most of the variation in psychiatric disorders is highly individualized and at the highest level of resolution (e.g. in terms of whole-brain voxel-level patterns of structural differences) does not provide compelling evidence for the existence of clusters.

---

## 7.6 Outlook and Challenges

There is a pervasive assumption that the optimal way to parse heterogeneity in psychiatric disorders is to partition the phenotype into subtypes. This assumption is effectively a recapitulation of the case-control approach and remains an implicit element of initiatives such as RDoC and ROAMER (Insel et al. 2010). Indeed, a criticism that has been leveled at RDoC is that it is in effect simply a new way to perform subtyping (Weinberger and Goldberg 2014). The subtyping approach has been successful in many other areas of medicine; for example, it has revolutionized oncology (Kalia 2015). However, we argue that it may not be optimal for psychiatric disorders. In psychiatry, few symptoms are unique to a single disorder and there are hundreds of genetic polymorphisms associated with most psychiatric disorders, all having small effect sizes and converging on similar symptoms (e.g. Betancur 2011; Ripke et al. 2014). Therefore, we argue that it may be unreasonable to expect cleanly separable subtypes for most disorders and alternative conceptual models may be more appropriate. One possibility is a ‘watershed’ model, which likens the pathophysiological process to a river system where many causative factors of small effect (e.g. genetic polymorphisms or environmental factors) begin as ‘tributaries’ and aggregate as they flow ‘downstream’ finding full expression in the syndromic expression of the disorder, akin to a river delta (Cannon 2016). Importantly, the

watershed model does not necessarily imply that subtypes will be evident in the data.

We have reviewed elsewhere the extensive literature aiming to partition psychiatric disorders on the basis of symptoms and psychometric variables, where we noted that this approach has still not converged on a consistent set of subtypes despite considerable effort (Marquand et al. 2016b). Here, we have focused on attempts to find biological subtypes or biotypes of psychiatric disorders. Whilst the studies we have reviewed suggest that this may be possible, none of these have been completely replicated at the present time and the degree of external validation of the derived subtypes is modest. More importantly, it is important to recognize that all the biotyping studies we have reviewed employed clustering algorithms, which always yield a result. In other words, they will return a specified set of clusters, regardless of whether the data support clusters. In general, there is no universal metric to determine the ‘optimal’ number of clusters or to adjudicate between different clustering algorithms for a given dataset, and as a consequence a proliferation of various metrics have been proposed (Marquand et al. 2016b). Unlike supervised learning, where there is a clear measure of model quality (i.e. the accuracy with which new samples can be predicted), unsupervised learning models can be compared in many different ways (e.g. cluster separability, reproducibility or external validation accuracy) and it is usually not clear which is ‘optimal’. Therefore, the final decision as to the ‘best’ clustering solution or algorithm often remains largely a matter of taste (Hastie et al. 2009). Moreover, most assessment metrics routinely used in practice are relative in the sense that they compare prospective clustering solutions with one another, but do not test the ‘null’ hypothesis that there are in fact clusters in the data. Various methods have been proposed that can be used to test whether clusters are ‘really there’ (Liu et al. 2008) and to compare the suitability of continuous, categorical and hybrid models for the data at hand (see Miettunen et al. 2016 for an overview). However, these approaches are currently underutilised for this purpose in psychiatry.

In line with this, it has been suggested that the biotypes reported by Clementz et al. (2016) may be better explained by a continuous dimensional representation relative to categorical subtypes (Barch 2017). We suggest here that the depression biotypes presented by Drysdale et al. (2017) may also reflect an imposed discretization of underlying continuous variation see Dinga et al. 2018 for further details <https://www.biorxiv.org/content/early/2018/09/14/416321>. In this study, biotypes were derived by training a clustering algorithm on two orthogonal mappings between brain connectivity and symptoms based on continuous subject loadings derived from canonical correlation analysis (Fig. 7.2a). Following cluster estimation, the authors increased the distinctiveness of their clusters by excluding ambiguous samples. Without this post-processing step, it becomes apparent that the evidence for the existence of clusters is equivocal in that the variation in the data could equally well be explained with two continuous axes (Fig. 7.2b). We emphasize that this does not imply that the findings reported are not biologically or clinically relevant, rather that the use of clustering algorithm imposes a categorical structure on the data that may not be optimal.

We reviewed several alternative methods to stratify psychiatric disorders. Whilst many of them are based on the same rationale as clustering approaches in that the phenotype can be split in to biotypes (Varol et al. 2017; Dong et al. 2016), these have features that ameliorate some of the problems inherent in applying ‘off the shelf’ algorithms to biometric data. For example, a common feature of many of these approaches (e.g. Marquand et al. 2016a; Varol et al. 2017; Dong et al. 2016) is that they break the symmetry inherent in the case-control and clustering approaches in the sense that they regard the disease cohort differently to the healthy cohort. This can be advantageous for stratifying psychiatric disorders because it allows the algorithm to focus on the manner in which patients deviate from a healthy pattern. It is especially beneficial in contexts where the clustering is performed on the basis of potentially high dimensional biological data because it means the clustering algorithm is less likely to detect nuisance variation that is of greater magnitude than disease-related effects (e.g. due to age or site).

Amongst the various methodological approaches we have reviewed, only a few methods are agnostic to the presence or absence of subtypes in the data (Miettunen et al. 2016; Marquand et al. 2016a; Mourao-Miranda et al. 2011; Zhang et al. 2016). Normative modeling is one promising example and whilst normative modelling can be used to derive features useful for clustering, its principal aim is to derive statistical estimates of deviation for each individual subject so that each subject can be compared to the normative or reference pattern. Another advantage of normative modelling is that it aims to estimate a supervised mapping and can therefore focus on the particular axes of variation (for example, the variation associated with a particular cognitive domain). Clearly, the development of alternative methods for stratifying the psychiatric phenotype are urgently needed.

As we briefly noted above, a major challenge for all methods is adequately and automatically dealing with artefacts in clinical datasets. There are many known sources of nuisance variance that are known to influence biological data and it is often the case that nuisance variation can be orders of magnitude greater than clinically relevant variation. This is particularly problematic because most stratification is performed in an unsupervised manner. A well-known example is head motion, which is widely acknowledged as a substantial challenge in fMRI studies (Van Dijk et al. 2012), and it is often the case that (in expectation) clinical groups move either more (e.g. ADHD) or less (e.g. depression) than healthy participants. These problems are compounded in large data cohorts, where data are often derived from multiple study sites, following different protocols. Moreover, nuisance variation often overlaps with clinically-relevant variation because important clinical or demographic variables are often not matched across study sites. Therefore finding techniques to deal with this optimally is a substantial ongoing challenge (Rao et al. 2017). One notable method that tackles this problem explicitly is the approach proposed by (Dong et al. 2016), which allows covariates such as age and sex to be specified so that the transformations estimated by the method take those into account.

## 7.7 Conclusions

In this chapter, we have reviewed literature aiming to use biological measures and big data cohorts to stratify psychiatric disorders. Whilst progress has clearly been made, there are major challenges for the field to overcome if we are to bring psychiatry closer towards precision medicine. We have argued that a widespread fixation on finding case-control type differences by partitioning the psychiatric phenotype into sharply defined clusters has impeded progress. Whilst successful in other areas of medicine, we argue that the complex multifactorial causes of psychiatric disorders combined with considerable overlap of symptoms across disorders mean that the biotyping approach may not be optimal in psychiatry. Currently only a few theoretical models have been proposed that do not assume the existence of clusters in the data (e.g. the ‘watershed’ model of Cannon 2016) and few analysis methods have been proposed that can fractionate psychiatric phenotypes without imposing clusters on the data. Alternative approaches are therefore urgently needed. Finally, we note that replication remains a major challenge for all methods. In line with the larger literature aiming to stratify psychiatric disorders (Marquand et al. 2016b; Schnack 2018), the studies reviewed here have—at best—performed a modicum of external validation, usually on the same cohort. At the time of writing, none of the studies we have reviewed in this chapter have been fully replicated to the degree that includes all steps in the analysis. This therefore remains an urgent priority.

---

## References

- Barch DM (2017) Biotypes: promise and pitfalls. *Biol Psychiatry* 82:2–3
- Bedi G, Carillo F, Cecchi G, Sezak GF, Sigman M, Mota N, Ribeiro S, Javitt DC, Copelli M, Corcoran CM (2015) Automated analysis of free speech predicts psychosis onset in high-risk youths. *Schizophrenia* 1:15030
- Beirlant J, Goegebeur Y, Teugels J, Segers J (2004) *Statistics of extremes: theory and applications*. Wiley, Sussex
- Betancur C (2011) Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res* 1380:42–77
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Boser B, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on computational learning theory*, vol 5, pp 144–152
- Brodersen KH, Deserno L, Schlagenhaut F, Lin Z, Penny WD, Buhmann JM, Stephan KE (2014) Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clin* 4:98–111
- Cannon TD (2016) Deciphering the genetic complexity of schizophrenia. *JAMA Psychiat* 73:5–6
- Clementz BA, Sweeney JA, Hamm JP, Ivleva EI, Ethridge LE, Pearlson GD, Keshavan MS, Tamminga CA (2016) Identification of distinct psychosis biotypes using brain-based biomarkers. *Am J Psychiatry* 173:373–384
- Costa Dias TG, Iyer SP, Carpenter SD, Cary RP, Wilson VB, Mitchell SH, Nigg JT, Fair DA (2015) Characterizing heterogeneity in children with and without ADHD based on reward system connectivity. *Dev Cogn Neurosci* 11:155–174
- Dong AY, Honnorat N, Gaonkar B, Davatzikos C (2016) CHIMERA: clustering of heterogeneous disease effects via distribution matching of imaging patterns. *IEEE Trans Med Imaging* 35:612–621

- Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, Fetcho RN, Zebley B, Oathes DJ, Etkin A, Schatzberg AF, Sudheimer K, Keller J, Mayberg HS, Gunning FM, Alexopoulos GS, Fox MD, Pascual-Leone A, Voss HU, Casey BJ, Dubin MJ, Liston C (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med* 23:28–38
- Gates KM, Molenaar PCM, Iyer SP, Nigg JT, Fair DA (2014) Organizing heterogeneous samples using community detection of GIMME-derived resting state functional networks. *PLoS One* 9(3):e91322
- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning*. Springer, New York
- Honnorat J, Dong A, Meizenzahl-Lechner E, Koutsoleris N, Davatzikos C (2018) Neuroanatomical heterogeneity of schizophrenia revealed by semi-supervised machine learning methods. In press
- Insel TR (2014) Mental disorders in childhood shifting the focus from behavioral symptoms to neurodevelopmental trajectories. *JAMA* 311:1727–1728
- Insel TR, Cuthbert BN (2015) Brain disorders? Precisely. *Science* 348:499–500
- Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, Sanislow C, Wang P (2010) Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* 167:748–751
- Ivleva EI, Clementz BA, Dutcher AM, Arnold SJM, Jeon-Slaughter H, Aslan S, Witte B, Poudyal G, Lu H, Meda SA, Pearlson GD, Sweeney JA, Keshavan MS, Tamminga CA (2017) Brain structure biomarkers in the psychosis biotypes: findings from the bipolar-schizophrenia network for intermediate phenotypes. *Biol Psychiatry* 82:26–39
- Kalia M (2015) Biomarkers for personalized oncology: recent advances and future challenges. *Metabolism* 64:S16–S21
- Kapur S, Phillips AG, Insel TR (2012) Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry* 17:1174–1179
- Kriegel H-P, Kroeger P, Zimek A (2009) Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans Knowl Discov Data* 3:1–58
- Lamers F, Vogelzangs N, Merikangas KR, De Jonge P, Beekman ATF, Penninx BWJH (2013) Evidence for a differential role of HPA-axis function, inflammation and metabolic syndrome in melancholic versus atypical depression. *Mol Psychiatry* 18:692–699
- Lecun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444
- Liu Y, Hayes DN, Nobel A, Marron JS (2008) Statistical significance of clustering for high-dimension, low-sample size data. *J Am Stat Assoc* 103:1281–1293
- Marquand AF, Rezek I, Buitelaar J, Beckmann CF (2016a) Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol Psychiatry* 80:552–561
- Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF (2016b) Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biol Psychiatry Cogn Neurosci Neuroimaging* 1:433–447
- Miettunen J, Nordstrom T, Kaakinen M, Ahmed AO (2016) Latent variable mixture modeling in psychiatric research—a review and application. *Psychol Med* 46:457–467
- Milaneschi Y, Lamers F, Peyrot WJ, Abdellaoui A, Willemsen G, Hottenga J-J, Jansen R, Mbarek H, Dehghan A, Lu C, CHARGE Inflammation Working Group, Boomsma DI, Penninx BWJH (2015) Polygenic dissection of major depression clinical heterogeneity. In press
- Miller KL, Alfaro-Almagro F, Bangarter NK, Thomas DL, Yacoub E, Xu JQ, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JLR, Griffanti L, Douaud G, Okell TW, Weale P, Dragonu J, Garratt S, Hudson S, Collins R, Jenkinson M, Matthews PM, Smith SM (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* 19:1523–1536
- Mirnezami R, Nicholson J, Darzi A (2012) Preparing for precision medicine. *N Engl J Med* 366:489–491
- Mourao-Miranda J, Haroon DR, Hahn T, Marquand AF, Williams SCR, Shawe-Taylor J, Brammer M (2011) Patient classification as an outlier detection problem: an application of the one-class support vector machine. *Neuroimage* 58:793–804

- Mwangi B, Matthews K, Steele JD (2012) Prediction of illness severity in patients with major depression using structural MR brain scans. *J Magn Reson Imaging* 35:64–71
- Rao A, Monteiro JM, Mourao-Miranda J, Alzheimers Dis I (2017) Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage* 150:23–49
- Rasmussen CE, Williams C (2006) Gaussian processes for machine learning. MIT Press, Cambridge
- Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu SA, Begemann M, Belliveau RA Jr, Bene J, Bergen SE, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Byerley W, Cahn W, Cai G, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Chan RCK, Chen RYL, Chen EYH, Cheng W, Cheung EFC, Chong SA, Cloninger CR, Cohen D, Cohen N, Cormican P, Craddock N, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, Demontis D, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Durmishi N, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedl M, Friedman JI, Fromer M, Genovese G, Georgieva L, Giegling I, Giusti-Rodriguez P, Godard S, Goldstein JI, Golimbet V, Gopal S, Gratten J, De Haan L, Hammer C, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Hollegaard MV, Hougaard DM, Ikeda M, Joa I et al (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511:421–427
- Ruiz FJR, Valera I, Blanco C, Perez-Cruz F (2014) Bayesian nonparametric comorbidity analysis of psychiatric disorders. *J Mach Learn Res* 15:1215–1247
- Sato JR, Rondina JM, Mourao-Miranda J (2012) Measuring abnormal brains: building normative rules in neuroimaging using one-class support vector machines. *Front Neurosci* 6:178
- Schnack H (2018) Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric disorders). *Schizophr Res*. In press
- Schumann G, Binder EB, Holte A, De Kloet ER, Oedegaard KJ, Robbins TW, Walker-Tilley TR, Bitter I, Brown VJ, Buitelaar J, Ciccocioppo R, Cools R, Escera C, Fleischhacker W, Flor H, Frith CD, Heinz A, Johnsen E, Kirschbaum C, Klingberg T, Lesch K-P, Lewis S, Maier W, Mann K, Martinot J-L, Meyer-Lindenberg A, Mueller CP, Mueller WE, Nutt DJ, Persico A, Perugi G, Pessiglione M, Preuss UW, Roiser JP, Rossini PM, Rybakowski JK, Sandi C, Stephan KE, Undurraga J, Vieta E, Van Der Wee N, Wykes T, Maria Haro J, Wittchen HU (2014) Stratified medicine for mental disorders. *Eur Neuropsychopharmacol* 24:5–50
- Sun H, Lui S, Yao L, Deng W, Xiao Y, Zhang W, Huang X, Hu J, Bi F, Li T, Sweeney JA, Gong Q (2015) Two patterns of white matter abnormalities in medication-naïve patients with first-episode schizophrenia revealed by diffusion tensor imaging and cluster analysis. *JAMA Psychiat* 72:678–686
- Torous J, Onnela JP, Keshavan M (2017) New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Transl Psychiatry* 7(3):e1053
- Van Dijk KRA, Sabuncu MR, Buckner RL (2012) The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* 59:431–438
- Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugarbil K, Consortium WU-MH (2013) The WU-minn human connectome project: an overview. *Neuroimage* 80:62–79
- Varol E, Sotiras A, Davatzikos C, Alzheimer's Disease Neuroimaging Initiative (2017) HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. *Neuroimage* 145:346–364
- Weinberger DR, Goldberg TE (2014) RDoCs redux. *World Psychiatry* 13:36–38
- Wolfers T, Buitelaar JK, Beckmann C, Franke B, Marquand AF (2015) From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev*. In press

- Wolfers T, Arenas AL, Onnink AMH, Dammers J, Hoogman M, Zwiers MP, Buitelaar JK, Franke B, Marquand AF, Beckmann CF (2017) Refinement by integration: aggregated effects of multimodal imaging markers on adult ADHD. *J Psychiatry Neurosci* 42:386–394
- Young J, Ashburner J, Ourselin S (2013) Wrapper methods to correct mislabelled training data. 3rd international workshop on pattern recognition in neuroimaging. IEEE, Philadelphia
- Zhang XM, Mormino EC, Sun NB, Sperling RA, Sabuncu MR, Yeo BT, Alzheimer's Disease Neuroimaging Initiative (2016) Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc Natl Acad Sci U S A* 113:E6535–E6544