Schizophrenia Research 214 (2019) 34–42

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases)

Hugo G. Schnack

Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht Univeristy, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 2 July 2017
Received in revised form 10 October 2017
Accepted 13 October 2017
Available online 1 November 2017

Keywords:
Machine learning
Heterogeneity
Prediction
Classification
Clustering

ABSTRACT

Psychiatric diseases are very heterogeneous both in clinical manifestation and etiology. With the recent rise of using machine learning techniques to attempt to diagnose and prognose these disorders, the issue of heterogeneity becomes increasingly important. With the growing interest in personalized medicine, it becomes even more important to not only classify someone as a patient with a certain disorder, its treatment needs a more precise definition of the underlying neurobiology, since different biological origins of the same disease may require (very) different treatments.

We review the possible contributions that machine learning techniques could make to explore the heterogeneous nature of psychiatric disorders with a focus on schizophrenia. First we will review how heterogeneity shows up and how machine learning, or multivariate pattern recognition methods in general, can be used to discover it. Secondly, we will discuss the possible uses of these techniques to attack heterogeneity, leading to improved predictions and understanding of the neurobiological background of the disorder.

1. Introduction

Psychiatric disorders such as schizophrenia and bipolar disorder are very heterogeneous. This heterogeneity complicates the investigation of the disorders and it comes in many different flavors. The forms that we are interested in from a scientific point of view are heterogeneity in the etiology on the one hand (Deng and Dean, 2013), reflected by heterogeneity found genetic profile (Liang and Greenwood, 2015), the neurobiological substrate (Brugger and Howes, 2017) and the clinical (Andreasen, 1995) manifestation on the other hand. Unraveling different mechanisms between biology/environment and clinical manifestation will enable us to make better diagnoses and prognoses, potentially leading to personalized medicine (Dazzan, 2014). Before going into detail about this 'true' heterogeneity, we briefly discuss what we call apparent heterogeneity (or apparent homogeneity). To study a disorder and the heterogeneity that comes with it, we need to carry out research, which inevitably concerns taking a –finite– sample from the population of patients. Factors that influence the homogeneity/heterogeneity of the disease, as it *appears* in the sample, come in two main flavors. The first is related to the inclusion criteria of the study, while the second form originates from technical factors.

The disease appears to be less or more heterogeneous depending on the inclusion criteria: Are males and females included, Caucasian and non-Caucasian subjects, with or without co-morbidities, and so on. If a study tries to control as many inclusion criteria as possible (e.g., a study on Caucasian males, aged 20 year, with a first psychosis, not using (of having used) any antipsychotics, without any comorbidities), one probably ends up with a quite homogeneous, but very small, sample. The power to find a biomarker for the disease would be very small. A potential biomarker found in such a study necessarily must have a large effect size – present (only) in patients with the characteristics of this sample. The general use of it is thus largely limited to an extremely small portion of 'patient space'. Small samples also have the risk of overfitting (see, e.g., Combrisson and Jerbi, 2015): the biomarker is not related to the disease itself but to unrelated properties of the subjects that accidently co-occur with the disease – in this sample.

On the other side of the spectrum are broad studies with hardly any restrictions for inclusion. Sample size will not be a problem for these studies but in this case as well a limited number of markers for the disease might be found (only the ones shared by most of the patients) and they might be non-specific: Because all patients need to display this effect it may not be limited to the specific disease, but be shared by patients with other diseases as well. The effect might, for instance, be secondary and be related to, e.g., changes in lifestyle. Potential biomarkers with large effect sizes (such as found in the homogeneous study) will not be found, because most patients' biological substrate is different from the one present in a homogeneous group.

For a detailed discussion of the impact of sample size on classification, see Schnack and Kahn (2016).

https://doi.org/10.1016/j.schres.2017.10.023
0920-9964/© 2017 Elsevier B.V. All rights reserved.

While these factors thus induce heterogeneity in the sample, most of them are known and are, in principle, under control of the researchers.

The second cause of apparent heterogeneity is of technical origin – but the mechanism by which it influences the properties of a sample is comparable. Measurements depend on who/what perform them. For clinical measures, this influence originates from the human rater who does the assessment and which assessment tool is used (for DSM-5, see Regier et al., 2013), while biological measures depend on the technical aspects of the measurement apparatus (e.g., MRI scanner; Kruggel et al., 2010) and protocols for acquisition and processing the data. While this heterogeneity is uninteresting and undesirable, it is also often unavoidable.

Summarizing, these factors lead to heterogeneity in the sample that is related to the setup of a study and can as such be controlled for (to certain extent). In that sense it is different from the 'true' heterogeneity that we will further discuss.

In this review we focus on machine learning methods to investigate the biological/clinical heterogeneity of schizophrenia, and psychiatric disorders in general. The discussion of possible (biological) origins of heterogeneity is not within the scope of this review. Nor is the machine learning methodology for making individual predictions in psychiatry, for which excellent reviews exist (see, e.g., Arbabshirani et al., 2017). Here we focus on the various forms of heterogeneity as we encounter it in studies (Section 2) and the way we can attack heterogeneity (Section 3). We will review the possible approaches and will provide references to (recent) studies employing specific approaches.

2. Encountering heterogeneity

2.1. Univariate and multivariate biomarkers

For disorders such as schizophrenia it is generally assumed that there are multiple pathways that lead to the disease (Deng and Dean, 2013). Single variables cannot capture this heterogeneity and show (undetectably) low effect sizes, but multivariate measures that incorporate the combined effects of many variables participating in a certain pathway could show large(r) effect sizes (Fig. 1). The important notion here is that, when studying differences at group-level, any effect size, however small, can be 'made' significant by taking larger samples. However, increasing the sample size does not improve the separability of the groups, since the non-overlap of the distributions is only related to the effect size and does not depend on *N*. Modest effect sizes lead to poor separations. Fig. 1 illustrates this point and shows effect sizes with corresponding separation accuracy for a number of univariate and multivariate biomarkers for schizophrenia from the literature.

2.2. Quantification of observed heterogeneity

Most diagnostic or prognostic prediction models aim to partition a sample into two or more classes, e.g. patients and controls, or different patient subgroups. The results of such classification models should be evaluated in light of the amount of heterogeneity present in the samples. Models from more heterogeneous samples are likely to produce less distinct classes of subjects, with corresponding lower prediction accuracy, while models from more homogeneous samples may show poorer generalizability (Schnack and Kahn, 2016). Quantification of heterogeneity is thus important when assessing a prediction model's quality. Within a sample, homogeneity clusters are usually measured by comparing within-cluster variation in subject measures (features) to between-cluster distance (in terms of average difference in measures). The primary use of such quality indices is to aid in determining the number of clusters that optimally split the sample. An example of a widely used indicator is the (average) silhouette score (Rousseeuw, 1987). In homogeneous samples of patients (and controls!) a two-cluster solution may be found optimal, while heterogeneous samples consisting of distinct subgroups are better split into more than two clusters. However, the heterogeneity encountered in psychiatric disorders tends to be diffuse, hindering the determination of the 'optimal' number of clusters.

Between samples, homogeneity can be assessed as the similarity of two (or more) prediction models: to which extent are the same predictors important for the classification. If subjects from two samples are represented by the same set of features (i.e., the same measures are available in both samples), each of the models trained to classify the subjects from a sample will do so by 'discovering' which of the features can perform this task. A feature being important for classification in one sample may be less important or not play a role at all in sample 2, and vice versa. The lists of discriminative features produced by each model can be compared and the proportion of discriminative features shared by the two classification models is a qualitative indicator of the (between-sample) disease homogeneity. For linear models these lists turn into weight vectors (**w**), indicating quantitatively each feature's influence, in relation to the other features', on having the disease. (Fig. 1, lower panel). In this case, between-sample homogeneity, defined as the proportion *f* of shared discriminative features, can be quantitatively related to the angle between the weight vectors: $f = \cos(\alpha)$ (Schnack and Kahn, 2016). An angle $\alpha = 0$ reflects identical models ($f = 1$) and increasing angles indicate less comparable models or, equivalently, increasing heterogeneity ($f < 1$). In a four-center machine learning study using neuroimaging data to separate first-episode schizophrenia patients from control subjects, Dluhoš et al. (2017) calculated $\cos(\alpha)$ between the weight vectors of the models built from the different sites. For the 6 different site-site combinations of models based on gray matter distributions, *f*-values between 0.06 and 0.35 were obtained [Suppl. Table 15 in Dluhoš et al., 2017], representing a *minimum* angle of 70° and suggesting a high level of heterogeneity. Many different factors may have played a role here. Lower *f*-values were obtained when making comparisons with smaller sample sizes (s3) and with a sample with longer illness duration (s2). While the low value of *f* in the latter comparison probably reveals true heterogeneity of the disease, the former may also be partly due to sampling effects. Furthermore, between-scanner differences may have played a role. These forms of apparent heterogeneity have been discussed in the Introduction.

To our knowledge the study by Dluhoš et al. (2017) is the only one to compare biomarker patterns obtained by machine learning models between different samples. From the literature it is currently not possible to study the heterogeneity of biomarker patterns, because of two reasons: (1) The wide variety in data acquisition and preprocessing techniques and in machine learning methodology; (2) The lack of detailed reports of the discriminative feature sets: Most studies have discussed the underlying discrimination patterns roughly, with little detail, or provided a few snapshot pictures or a list of only the most important features. An exception is Dluhoš et al. (2017), who made their models available as weight maps in the familiar nifti format.

In summary, measures such as silhouette score and *f*-value reflect heterogeneity within and between samples; they may be used to compare different classification solutions in the approaches discussed in Section 3.

3. Attacking heterogeneity

In this section we review and discuss several approaches to deal with heterogeneous samples, when using machine learning techniques for diagnosis or prognosis of psychiatric diseases.

3.1. Do not attempt to account for heterogeneity

Most published studies thus far have taken this approach (for schizophrenia, see, e.g. review by Kambeitz et al., 2015). In this case, a (linear) classification algorithm is applied to separate patients and controls, and one accepts a possible suboptimal performance: only the larger, shared, biomarkers (or main pattern) will be found and the prediction accuracy will be modest. (Fig. 2, upper left). This means that one will not fully

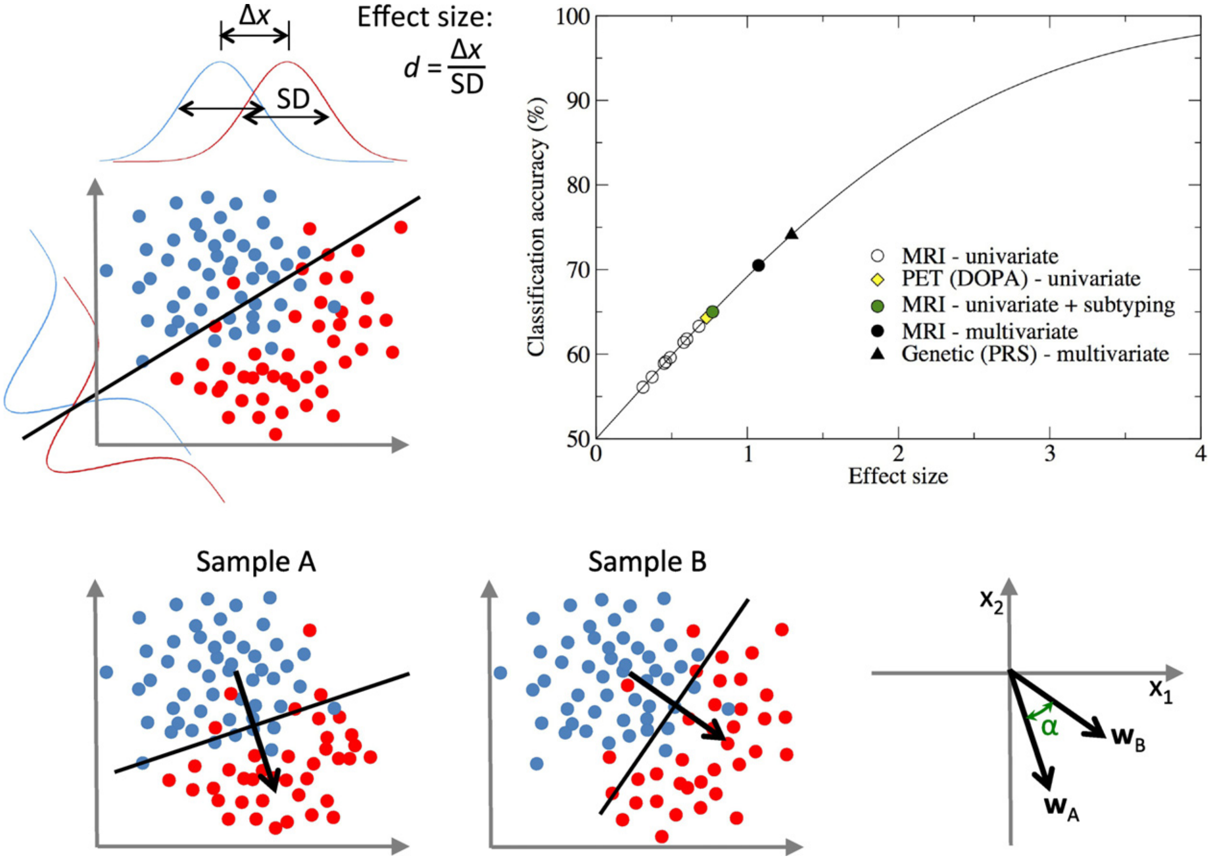

Fig. 1. (Top panel) Upper left: Distributions of a feature (e.g., ventricle volume) for the subjects of two classes, e.g., normal (blue) and patient (red). The effect size of the difference between the two distributions, Cohen's d, is defined as the ratio of the mean difference and the width (standard deviation) of the distributions. A univariate statistical test (e.g., a t-test) is used to determine whether or not there is a significant difference in this feature at group-level. The effect size is directly related to the non-overlap of the distributions: the classification accuracy when they are optimally split. However, while increasing the sample size will lower the detection limit of a significant difference (because $t \sim d \times \sqrt{N}$), it will not help in splitting patients and controls based on this variable: The overlap is related to the effect size itself and does not depend on N. Lower left: Multivariate biomarkers can yield larger effect sizes. The right panel shows the relationship between effect size d and non-overlap, represented by the separation accuracy (percentage correctly classified subjects) (line). The symbols represent effect sizes of studies comparing schizophrenia patients and controls together with their corresponding classification accuracy when the biomarkers would be used for classification. Open circles: several volumetric measures (including gray matter, thalamus, ventricles) from MRI meta-analyses (Haijma et al., 2013; van Erp et al., 2016); yellow diamond: meta-analytic average of dopamine synthesis capacity from PET (Howes et al., 2007); Green circle: adhesio interthalamica length from MRI (Takahashi et al., 2017); Black circle: multivariate gray matter density profile from MRI (Nieuwenhuis et al., 2012); Black triangle: polygenic risk score (PRS) (So and Sham, 2017). (Bottom panel) The left and middle graphs again show distributions of patients (red) and controls (blue) of two samples (A and B) in a two-dimensional feature space (spanned by features x_1, e.g., ventricle volume, and x_2, e.g., brain volume). The heterogeneous nature of the disease is illustrated by the fact that, in sample B, patients can be separated from controls based on having larger x_1 (ventricle volume) as well as smaller x_2 (brain volume), whereas in sample A, the separation is mainly based on having smaller x_2, with x_1 not playing a big role. The orientations of the separation lines (hyperplanes in higher-dimensional feature spaces) are defined by vectors orthogonal to the hyperplanes, the so-called weight vectors, $\mathbf{w}_A$ and $\mathbf{w}_B$ (right graph); the angle α between them is a measure of the difference in orientation between the hyperplanes and, thus, of the differences in importance (weight) of the features (x_1, x_2) between the models/samples. Here, $f = \cos(40^\circ) = 0.76$, which could be loosely related to the proportion of shared features lying between $1/2 = 0.5$ and $2/2 = 1$.

make use of the increased power of multivariate analysis (Fig. 1, bottom left). On the other hand, these whole-sample (linear) models will be quite robust, show little overfitting and may thus generalize better.

3.2. Treating heterogeneity (in)directly

One may assume that there is heterogeneity present in the sample that prevents simple (linear) separation. To solve this problem one could still try to separate patients and controls in one operation, but one has to allow for a more complex decision boundary between the classes, by adapting linear classifiers to deal with nonlinearity or choosing classifiers that are inherently nonlinear:

3.2.1. Transforming the data

Simple nonlinear relationships may be implemented by, e.g., adding quadratic transforms of the features to the feature set (Fig. 2, upper right). A disadvantage of this approach is that specific nonlinearities are modeled this way and that the number of features rapidly increases (in the case of quadratic transformations by a factor 2 or 3, depending on whether interaction terms are included). Only if knowledge about the possible mechanisms is available, one could implement the exact transformations necessary for describing these mechanisms. A more general way to implement nonlinear transformations is to use a kernel function, which models the similarity between data points. A frequently used kernel is the radial basis function (RBF) which has been applied in a schizophrenia outcome prediction study (Koutsouleris et al., 2016). A disadvantage is the risk of overfitting and the difficulty to interpret the models, i.e., understanding the relationship between features (predictors) and output.

The following approaches (3.2.2-3.2.4) make decisions based on the fact that heterogeneity in feature space can be described using combinations of AND and OR operators.

3.2.2. Artificial neural networks (ANNs)

If patients with different substrates can be described as clearly lying in different portions (polytopes) of feature space, bounded by clear hyperplanes, a combination of linear classifiers, each describing one hyperplane, could perform the classification. In this case, the individual classifiers are combined using AND/OR operators (Fig. 2, 2nd row). The multi-layer perceptron (MLP) is an example of such a construction of linear classifiers (perceptrons) and AND/OR operators. In principle, any heterogeneity can be modeled with this approach, and its generalization, artificial neural networks (ANNs). Each instance of an AND/OR situation is modeled by a node of the network. Different nodes are placed in one or more (hidden) layers. ANNs are very powerful and can virtually fit any relationship but at the cost of increasing the number of nodes. Larger numbers of input nodes (features) and/or nodes in hidden layers lead to a rapid increase of the number of parameters (weights) that need to be fitted. To avoid overfitting (i.e. adapting the model too much to the training data, causing poor generalization), very large samples are needed. This is especially a problem in medical setting, where patient data are often scarce. ANNs may arrive at local optima, thus many runs may be needed to find the global optimum. They are also difficult to interpret because of the complex flow of information through the nodes of the different (hidden) layers. Recently, new insights have led to the development of deep learning networks, which denotes networks with multiple layers to model different abstraction levels of learned features. This kind of networks seems to be very suitable for (natural) image recognition, but the field is rapidly evolving and one may speculate that it can be developed for use in (heterogeneity in) psychiatric diseases. First attempts have been made: Payan and Montana (2015) applied convolutional neural networks to structural MRI images to classify Alzheimer's disease in a sample of $N = 2265$ subjects.

3.2.3. Combining linear SVMs

While complex, nonlinear, modeling (Sections 3.2.1 and 3.2.2) may solve the classification, it does not provide any information about the underlying subgroups of patients that led to the heterogeneity. In (Varol et al., 2017), the authors propose to tackle the problem by combining linear classifiers (SVMs). Their approach will increase classification accuracy and subtype the pathology. Support vector machines (SVMs) are extended to a more general framework in order to do binary classification and subtype identification simultaneously. HYDRA is a non-linear semi-supervised machine learning algorithm that combines multiple (K) SVM classifiers to create a convex polytope that separates the healthy controls from the heterogeneous group of patients. (Fig. 2, 2nd row, right). The controls are assumed to form a homogeneous class. The degree of heterogeneity can be varied by choosing the number of hyperplanes. Each patient will be assigned to one of the hyperplanes to be separated from the controls. Patients that have been assigned to the same hyperplane

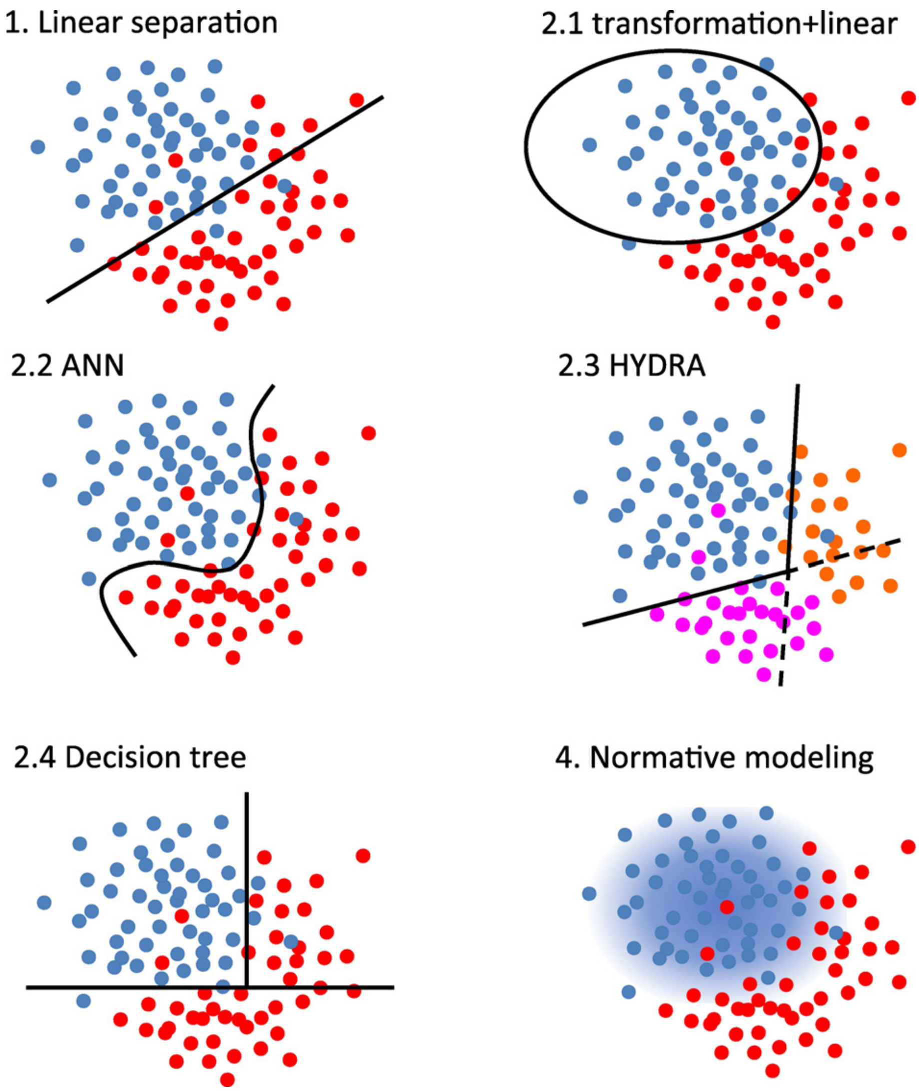

Fig. 2. Different approaches to attack heterogeneity. Patients are represented by red circles, controls by blue circles. Black lines represent the decision boundaries. See the text in Section 3 for descriptions of the methods. For the approaches 3.2.5–3.2.7, two feature spaces are shown: biological space in the left column, clinical space in the right column. Numbers 1 and 2 indicate two consecutive steps. Orange and pink circles represent different patient subtypes or clusters.

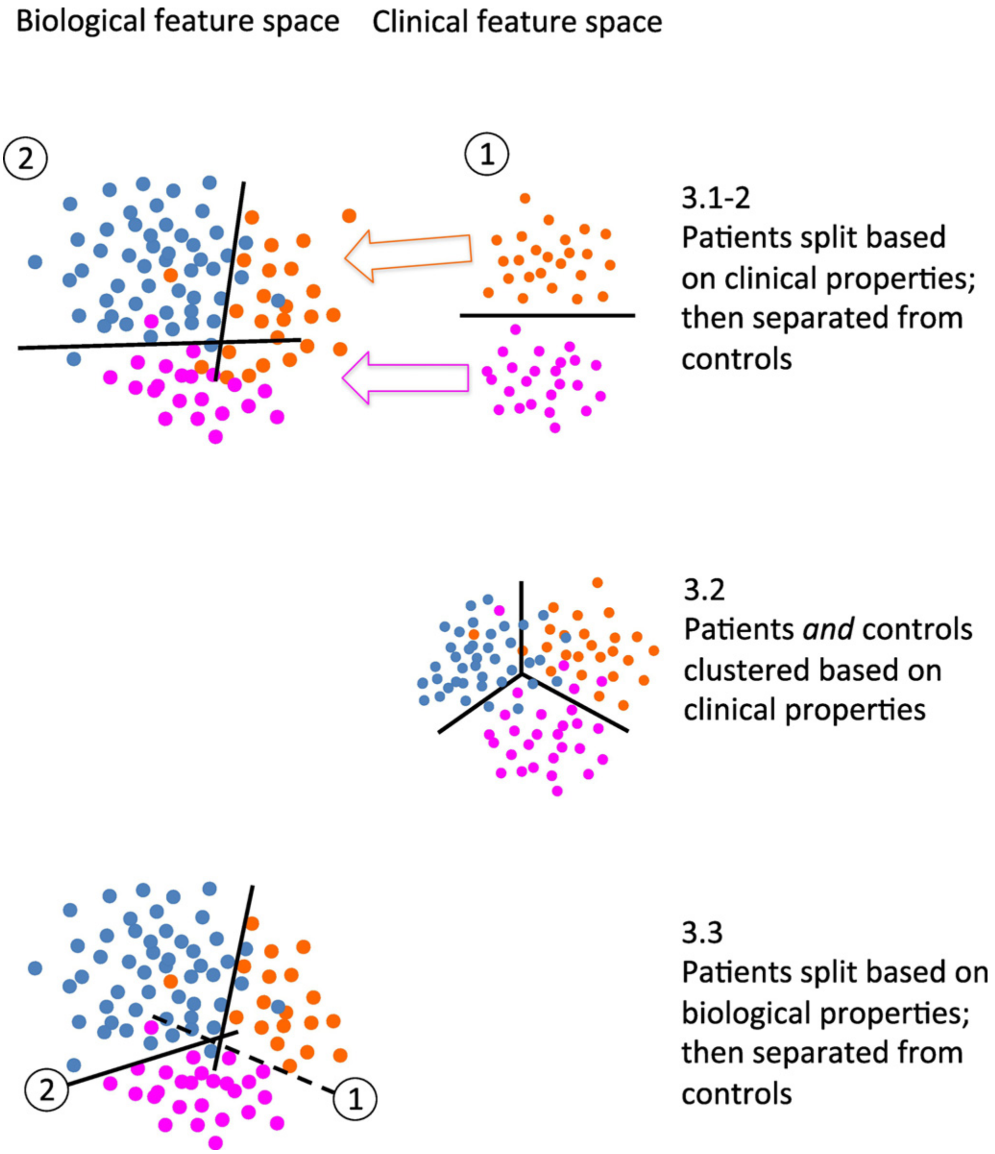
Biological feature space
Clinical feature space
2
1
3.1-2
Patients split based on clinical properties; then separated from controls
3.2
Patients *and* controls clustered based on clinical properties
2
1
3.3
Patients split based on biological properties; then separated from controls

Fig. 2 (*continued*).

are considered to be part of the same subgroup, indirectly rendering the problem also into a clustering task. When applied to a sample of patients with Alzheimer's disease and controls ($N = 300$), classifiers using $K > 1$ did not perform significantly better than the $K = 1$ classifiers. However, the other aim, to increase the margin, was achieved for $K > 1$, implicating, according to the authors, the algorithm had successfully found heterogeneous structures in the data. A three-cluster solution showed the highest stability. The subtyping is a secondary effect of finding the best discrimination between patients and controls, as opposed to clustering methods that primarily focus on finding the best division (see Section 3). Both approaches share the disadvantage that the number of clusters needs to be chosen, a disadvantage that is not present for ANNs. To our know knowledge, HYDRA has not yet been applied to schizophrenia data.

3.2.4. Hierarchical methods such as decision trees, random forests

This is another form of an AND/OR decision algorithm, where a subject's classification depends on the endpoint of which (OR) branch (sequence of ANDs) it ends up. Each decision in the tree (node) is based on a single variable. Consequently, where HYDRA can divide the feature space using any convex polytope and ANNs even in any possible form, trees partition feature space using hyper-cuboids oriented along the axes representing the original features. (Fig. 2, 3rd row). Decision boundaries formed by linear combinations of features (oblique hyperplanes) are thus not easily implemented by trees. However, decision trees are able to model any complex decision boundary, albeit sometimes at the cost of making the tree very large. Individual decision trees show large variance, which can be solved by taking ensembles of them: random forests. However, while simple trees are easy to interpret, larger trees and ensembles are, like ANNs, difficult to interpret.

While the above methods attack heterogeneity in order to improve classification accuracy, we do not learn much about heterogeneity from them. Patients and controls are treated as two (homogeneous or heterogeneous) distributions that needed to be separated with one model "in one run." The resulting models might be too complicated to be fitted reliably and/or too difficult to interpret. If, however, training such an algorithm would lead to a model with high, validated, prediction accuracy, one could be pragmatic and accept it as a clinically useful model. In psychiatry, however, prediction models do not have reached accuracies that are considered clinically relevant (First et al., 2012). Trying other directions may thus be useful, if these can provide us with insight in the heterogeneous structure of the disease:

Approaches that consist of a step to reduce the heterogeneity by stratifying the patients into subgroups, often followed by a classification step to separate each of the patient subgroups from the control group (3.2.5-3.2.7):

3.2.5. Subdivision of patients according to explicit, clinical, criteria

If one has a theory or expectation that a population of schizophrenia patients is made up of subgroups with different etiology (biomarkers), one can easily subdivide the group according to criteria defining these specific subgroups (Fig. 2 (cont.), 4th row).

For example, Takahashi et al. (2017), who used the PDS, a PANSS-derived identification tool, to split their sample of schizophrenia patients into the deficit subtype and non-deficit subtype. Looking for structural neuroimaging biomarkers for schizophrenia, they found differences in brain anatomy mostly between the deficit group and the other groups. The adhesio interthalamica was shorter in patients than in controls (Cohen's $d = 0.53$), but in the deficit group is was much shorter ($d = 0.77$) while in the non-deficit group the length difference was smaller ($d = 0.34$).

Czepielewski et al. (2017) followed a similar approach but investigated the influence of estimated premorbid crystallized IQ (ePMC-IQ) on structural brain abnormalities in schizophrenia. They found that patients with ePMC-IQ and cognitive impairments had a more extensive pattern of gray matter reductions than those with intact ePMC-IQ.

The potential gain in biomarker detection by homogenizing the patient group, as shown by both studies, is twofold: larger effect sizes are created and different biomarker – clinical presentation relationships can be resolved. The investigators applied univariate tests; it would be interesting to see if biomarkers with even higher effect sizes could be constructed by employing multivariate (machine learning) techniques.

Gould et al. (2014) split their patient group according to differences in cognition between patients. They defined two patient subtypes: 'Cognitive deficit' and 'cognitively spared'. The authors then built a classification model to separate schizophrenia patients and control subjects based on structural MRI images. The authors separated patients and controls with ~70% accuracy, but did not find a significant increase in classification accuracy between the whole-sample model and the models using distinct cognitive subtypes. They further verified this by cross-applying the cognitive-subtype models to the other subtype, resulting in about the same accuracy. The conclusion that the discriminative sMRI patterns largely overlap between the two cognitive subtypes was further confirmed by the high correlation ($r = 0.67$, roughly comparable to the quantity $f = \cos(\alpha)$ in Section 2) between the two weight maps.

Apart from these well-defined subtypes by theoretical assumptions or clinical evidence, there are probably more subtypes that are based on combinations of patient characteristics in a way that is not clear beforehand. Instead of applying explicit split criteria to one or more patient characteristics, one can let an algorithm do the subgrouping, which can be done from different directions (Sections 3.2.6 and 3.2.7):

3.2.6. Clustering based on clinical/demographic variables

In this approach, an algorithm tries to find patterns in the (clinical) data that define patient subtypes. It can be used if there is no clear way to split the patient group (see Section 3.2.5.), or if one assumes different subtypes are determined by a more complex, multivariate, pattern. Put simply, the task is to divide a heterogeneous population into more homogeneous subgroups. This process is called clustering, a unsupervised form of machine learning, and there are many possible algorithms that can do it.

Dickinson et al. (2017) applied a 2-step clustering algorithm to divide large sample of schizophrenia patients ($N = 549$) into "deficit", "distress" and "low symptom" subgroups, based on two PANSS composite scores. The patients in the "deficit" and "distress" groups showed both shared and distinct deviations in clinical and behavioral variables when compared to the "low symptom" group. A subset ($N = 182$) of the patients underwent fMRI scanning with the N-back working memory paradigm. The "deficit" subgroup showed a different activation profile than the two other groups, indicating a possible different etiology (Fig. 2 (cont.), 4th row).

A comparable approach, but ignoring the patient/control distinction, was carried out by Derks et al. (2012). In this study, latent class analysis was combined with factor analysis of clinical (CASH) data to obtain a split of a very large sample ($N = 4286$) of patients with psychosis, their relatives and unrelated control subjects into 7 subgroups, or 'classes': 'Kraepelinian schizophrenia', 'affective psychosis', 'manic-depression', 'deficit nonpsychosis', 'depression', 'healthy', and 'no symptoms'. Their approach was cross-diagnostic and included healthy control subjects and patients with psychosis (schizophrenia, schizoaffective disorder, bipolar disorder) who were redistributed over the 7 latent classes. Class membership showed associations with cognition (WAIS IQ) and functioning (CAN). While this approach does not focus on separating patients from controls, the authors state that it may improve the effectiveness of, e.g., finding genes related to symptom profiles, since clinical heterogeneity is reduced not only in patients but also in non-patients. (Fig. 2 (cont.), 5th row).

3.2.7. Clustering based on biological features

Clementz et al. (2016) applied a two-step clustering approach to divide a large sample of patients across the psychosis spectrum ($N = 711$; schizophrenia, schizoaffective, bipolar with psychosis), their relatives ($N = 883$) and controls ($N = 278$) into three classes of different 'biotypes'. First, the patient sample was used to create composite scores of biomarkers of different modalities (EEG, saccades, BACS) using PCA. A clustering step was then applied to the composite score data to divide the patients into three biotypes, each representing a different composite pattern. Interestingly, biotype did not correlate with the clinical diagnosis; however, patients of two, but not the third, biotypes differed significantly on cognitive control from healthy controls. Biotypes 1 and 2 (but not 3) also displayed comparable, extensive brain abnormalities as compared to the healthy group. Apart from, again, increasing effect sizes, these results demonstrate the benefit of looking across clinically (DSM) defined boundaries, the use of which has been recently introduced (Insel et al., 2010).

The method of brain subtyping (Dwyer et al., 2017), using fuzzy c-means clustering, splits a patient group into two or more subgroups, each with a different so-called 'braintype'. In a sample of $N = 145$ schizophrenia patients and healthy control subjects, an optimal separation into two schizophrenia subgroups was found. After this subdivision a (linear) classifier (SVM) was applied to separate controls from each subgroup. The classification accuracy increased from 68.5% without subtyping to 73 and 79% for the two subtypes, respectively. While these substrate-based divisions need not be related to clinical patient types, Dwyer et al. showed that they did in fact correlate with age, illness duration, sex and differential patterns of psychotic symptoms. (Fig. 2 (cont.), 6th row).

An advantage of these clustering/subtyping approaches is that the resulting classification models are simpler and point to clearer aberrant properties of patients. A drawback of these methods is that, although statistical measures can guide this process, the number of clusters needs to be chosen – a choice that may have substantial influence on the resulting subtypes and the etiological meanings that are attached to them.

Most approaches discussed thus far focus on the separation of patients and controls (or, more generally, dividing individuals into distinct classes) and are called discriminative modeling. A completely different approach is to refrain from this patient-control discrimination and model both groups separately as two independent distributions. This is called generative modeling, and it leads to estimates of probability distributions for each class. New subjects can be attributed to one of the classes based on the membership probabilities. See Libbrecht and Noble (2015) for a discussion in the domain of genetics. The benefit of generative modeling is that it provides a full description of the two classes, rather than focusing on the differences between the two classes. For our purpose, however, it may not work that well, exactly because of the heterogeneous nature of the patient group; an accurate and reliable description of this class might be difficult, and may probably require unrealistic large samples.

A variation of generative modeling is normative modeling, where only the 'normal' distribution is modeled, and subjects ('patients') may be detected as outliers:

3.2.8. Normative modeling

In this approach (Marquand, Wolfers et al., 2016), the goal is to estimate the probability distribution of healthy individuals (Fig. 2, bottom row). This distribution can have any shape, but it can be thought of as a multi-dimensional Gaussian distribution, with most people being located in its center, becoming more diluted (smaller probability) farther away from the center. Advantages are twofold: from a scientific point of view, this distribution modeling connects very well with the principle of normal variation due to multiple factors influencing a biomarker or phenotype (trait); from a clinical (and technical) point of view, there is no need to either invoke complex separation borders or make choices about how to split up the patients. After the normative modeling, outliers, subjects who have very small probability to be part of the normative distribution, can be detected and being labeled as not-normal, or 'patient'. Viewing a patient as an extreme case of the normal variation, need not be an advantage per se, since there might very well be a distinct 'disease class'. Furthermore, to 'detect' the patients, one still has to choose a probability threshold. While interesting and promising, some of the method's properties need to be further addressed. The main question is: who is included to be part of the normal group? Marquand, Rezek et al. (2016) used a healthy cohort ($N = 491$) to map the relationship between behavior (trait impulsivity) and biology (fMRI-derived reward-related brain activity) and related extreme values to ADHD symptoms. If people with subclinical symptoms are included in the normative distribution, again an artificial threshold has to be set. (Note that this issue also plays a role in discriminative modeling.). Including subjects not having the disease of interest, but other diseases, may influence the estimated distribution, while if patients with the disease of interest are included in the modeling, incidence rates need to be carefully taken into account. Large samples are necessary to lower the influence of these inclusions, and, indeed, the authors state that large cohorts of healthy ('normative') subjects are needed to capture 'all' variation and get a good description of the distribution.

Normative modeling may be very well applied to high-risk groups, where it could be clinically very useful to detect as early as possible potential deviations from normal development. In clinical situations however, when it may be clear to a psychiatrist that a person has some disorder – but finds it difficult to determine which disorder, the method cannot aid in the differential diagnosis, necessary for determining the right treatment. In such situations, discriminative modeling may be more useful (see, for example, the case of discriminating between schizophrenia and bipolar patients, a clinically very relevant task in first episode patients (Schnack et al., 2014)).

For prognostic applications, such as the problem of predicting outcome, the use of normative modeling may require modifications, since there is no 'normal' group here, only a heterogeneous patient group.

Marquand, Wolfers et al. (2016) gives an excellent overview of the various methods discussed in Sections 3.2.7 and 3.2.8 along with their technical details.

Finally, one could take a somewhat different route:

3.2.9. Multi-modal prediction models

If a given feature set turns out to be suboptimal for the task, i.e., if part of the patients is and part is not (even nonlinearly) separable from controls, it might seem that a homogeneous biological substrate (based on the current selection of features) gives rise to a heterogeneous clinical presentation. Apparently, we miss something: there are 'hidden variables' interacting with the current set of features, splitting the apparent homogeneous group of subjects into two (or more) groups with different disease status. An example of this is gene-environment interaction, where subjects with a certain genotype will get the disease if they have experienced a specific environmental factor. Adding features from different modality (e.g., environment) could solve this problem of poor separability.

Combining features from different sources can be done by simply putting all features together, but another option is to use multi-kernel learning, in which sub-models per modality are combined into a full model (see, e.g., Squarcina et al., 2017). The use of multiple kernels is especially useful in cases where data types are very different and when there is a large unbalance between numbers of features of the different modalities.

3.3. Heterogeneity in disease course/outcome: longitudinal studies and using machine learning for prognosis

The approaches discussed thus far mainly focus on partitioning the heterogeneous group of patients in a sample into more homogeneous subgroups. This subdividing was driven by either the biological or the clinical properties of the subjects – of the moment. No link with the future state (biological or clinical) of the patient was made. As such these approaches are thus mainly of interest from a scientifically point of view: they can improve the understanding of the etiology of the disorder. The clinical use of separating patients from controls, however, is limited. Of more interest is the separation of different clinical groups (e.g., schizophrenia and bipolar disorder (Schnack et al., 2014)), and the question what the disease course will be; what will be the outcome, which medication will work best for an individual patient? To answer these questions one need to perform longitudinal studies in which patients are followed through time. Machine learning models can be used to find patterns in baseline data of patients that are related to the course of the disease. Heterogeneity shows up again in these problems, since apparently comparable patients may show very different outcomes. It is probably this heterogeneity that is responsible for the modest prediction accuracies of the few outcome prediction studies that have been published to date: Mourao-Miranda et al. (2012) reached 69% prediction accuracy using baseline structural MRI data to predict illness course. A multi-center study (Nieuwenhuis et al., 2017) could not replicate this result: most probably heterogeneity both in the clinical description of illness course and in the underlying brain substrate (apart from MRI-technical differences) certainly played a role here. In an attempt to reduce the heterogeneity, the authors split the sample in males and females; the accuracy of the male-only models increased in some of the participating centers. A limitation of the study was the relative small sample size ($N = 212$) as compared to the number of centers (5), also making it difficult to split the sample in subgroups for improved analysis. Recently, a multi-center study predicting 1-year functional outcome (good versus poor) obtained 71% prediction accuracy, based on clinical/demographic variables (Koutsouleris et al., 2016). Although this study had an even greater unbalance between sample size ($N = 334$) and number of sites (44), the multi-center study was set-up using strict inclusion criteria, harmonized assessments, and did not use measures that are influenced by technical differences such as MRI.

Literature on this topic is too limited to draw conclusions about the feasibility of predicting outcome (Dazzan et al., 2015). New studies hopefully will shed light on which data source(s) can be best used to predict illness course and outcome. A subtle difference here is the notion that illness course is the description of the states of a patient during a period of time, while outcome is a measure of the patient's state at a certain point in the future – the endpoint of the course. One could speculate that more accurate predictions of outcome could be made if 'static' baseline data would be enriched with directional data of the early changes due to treatment. Inclusion of data reflecting the patient's state at multiple (dense) time points during the early phase of (treatment of) the disease could thus improve the predictions of outcome later on.

4. Conclusions

When reviewing the role of heterogeneity in machine learning studies in schizophrenia and other psychiatric diseases we have seen that: (1) heterogeneity indeed shows up, evidenced by a) differences in biomarker patterns encountered in machine learning studies on comparable clinical groups (Section 2) and b) improved classification accuracy when patients groups were first homogenized (Sections 3.2.5-3.2.7); (2) there are many possible approaches to attack heterogeneity (Sections 3.2–3.2.9). We encountered a variety of flavors of clustering and separating: Methods that first 'stratify' patients, reducing the heterogeneity in the sample, and then separate them from controls; and methods that cluster patients and controls simultaneously. There are methods that perform clustering (subtyping) in biological feature space and methods that subdivide patients in clinical feature space. Some variations are missing; the most interesting probably the clustering based on biological and clinical features combined. Rather than splitting/clustering in one space and testing/verifying whether the obtained subgroups make sense in the other space, the combined approach should be able to deliver subtypes of the disease that are well defined in both domains. On the other hand it should be noted that 'validating' a subtyping obtained in one domain by showing its correlation with certain features in the other domain has its limitations. While a positive finding may be used as support for the subtyping, a negative finding can be 'explained' as showing that the clinical classifications (DSM diagnoses) have limited meaning when searching for neurobiological substrates of these disorders.

A direct comparison of the different approaches is lacking in almost all studies. Some studies compared different machine learning algorithms (viz. linear and various nonlinear algorithms; Sections 3.2.1–3.2.4) in one sample. On the other hand, the approaches that explicitly deal with heterogeneity in patients (Sections 3.2.5–3.2.7) have only been tested against the reference: the proposed method versus the 'homogeneous' method (all patients versus all controls), but not with other heterogeneity reduction methods. Apart from increasing classification accuracy by these methods, other factors may also influence researchers' or clinicians' choice for a method. These factors include, e.g., the importance one attaches to model interpretability and practical issues, such as data availability or sample size.

A more fundamental point is whether or not disease heterogeneity is categorical or continuous. While it is most likely that there are many different pathways to disease that are not fundamentally different, giving rise thus to marginally different biological substrates which will show up as a heterogeneous continuum (or continuous heterogeneity), there might also exist a (limited) number of fundamentally different families of pathways. For this kind of heterogeneity a fruitful approach may be to combine different methods, e.g., subtyping (Sections 3.2.5-3.2.7) the different families together with modeling heterogeneity continuously by using techniques comparable to those employed in normative modeling (Section 3.2.8).

To further map heterogeneity it is crucial that subtyping and classification models are compared against each other: modeling tools and model parameters should be made publically available, so that other researchers can examine them and compare them with their own models. A few studies have already published their models (Koutsouleris et al., 2016; NeuroMiner, 2017) and Dluhoš et al., 2017. Prediction models should also be tested in independent samples, the results of which indicate relative heterogeneity between the training and testing samples. A complicating factor for model exchange is the differences in available measurements. MRI is often not available, and fMRI can have been acquired using different paradigms. Clinical, cognitive, demographic sources of data may show even larger variety.

Another point that is clear, is that current diagnostic (see, e.g. Kambeitz et al., 2015) and prognostic (see Section 3.3) models are not good enough for clinical applications. While models based on small samples may reach (unvalidated) accuracies of 80% or higher, larger studies show accuracies of about 70% (Schnack and Kahn, 2016). While measurement noise, sampling effects (potentially be solved by increase sample size) and imperfect gold standards (Regier et al., 2013) will always lower the practically highest possible accuracy (Schnack and Kahn, 2016), machine learning approaches taking care of clinical and/or biological heterogeneity will contribute to better models. However, improvements are also expected from combining different data modalities into one model (Section 3.2.9). This can be 'within-domain' multimodality, such as combined structural and functional MRI, and 'cross-domain' multimodality, such as combining MRI, genetics and demographic data.

Finally, large samples are needed to build accurate and reliable models, with enough cases for training and testing, capturing as much of the variation of the disorder as possible, and allowing subdividing the sample into subgroups. Multicenter studies are excellent for this purpose. Quite a few of the studies reviewed were multicenter. While including patients from multiple centers does not only increase sample size, it also almost automatically increases the biological heterogeneity, e.g., by including subjects with different genetic background. Models can be built either mega-analytically (data of all subjects to be collected at one site, see, e.g., Orban et al., 2017) or meta-analytically (sub-models built per site, see Dluhoš et al., 2017). Although the size of studies is an important factor, their design also needs attention. Naturalistic cohorts and combinations of them are very suitable for further exploration of (bio)markers for the disease and its course, because they can easily be made very large, while capturing as much heterogeneity as possible. On the other hand, randomized clinical trials with different treatments are necessary. Such studies allow for stratification of patients with respect to their responses to different treatments, so that machine learning can be used to discover the (neurobiological) substrate related to these differences in response. Prediction models can be built to predict the responses to different treatments, leading to targeted, or personalized, medicine (DeLisi and Fleischhacker, 2016; Keshavan et al., 2017). These models could subsequently be validated (retrospectively) in naturalistic cohorts.

In conclusion, machine learning has shown to be a promising technique for detecting heterogeneity in schizophrenia (and other psychiatric diseases) and reducing it to improve understanding its etiology and making predictions for individual patients possible. To further enhance its efficacy, future studies should be multi-center, using multi-modal predictor data, allowing for modeling more complex biomarker – clinical presentation relationships.

Conflicts of interest

Hugo Schnack reports no conflicts of interest.

Contributors

Hugo Schnack carried out the research for this review and wrote the manuscript.

Role of the funding source

Not applicable.

Acknowledgments

We thank Suzan Stempher for useful discussions.

References

Andreasen, N.C., 1995. Symptoms, signs, and diagnosis of schizophrenia. Lancet 346 (8973), 477–481.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145 (Pt B), 137–165.

Brugger, S.P., Howes, O.D., 2017. Heterogeneity and homogeneity of regional brain structure in schizophrenia: a meta-analysis. JAMA Psychiat. Sep 27. https://doi.org/10.1001/jamapsychiatry.2017.2663.

Clementz, B.A., Sweeney, J.A., Hamm, J.P., Ivleva, E.I., Ethridge, L.E., Pearlson, G.D., Keshavan, M.S., Tamminga, C.A., 2016. Identification of distinct psychosis biotypes using brain-based biomarkers. Am. J. Psychiatry 173 (4), 373–384.

Combrisson, E., Jerbi, K., 2015. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J. Neurosci. Methods 250, 126–136.

Czepielewski, L.S., Wang, L., Gama, C.S., Barch, D.M., 2017. The relationship of intellectual functioning and cognitive performance to brain structure in schizophrenia. Schizophr. Bull. 43 (2), 355–364.

Dazzan, P., 2014. Neuroimaging biomarkers to predict treatment response in schizophrenia: the end of 30 years of solitude? Dialogues Clin. Neurosci. 16 (4), 491–503.

Dazzan, P., Arango, C., Fleischacker, W., Galderisi, S., Glenthøj, B., Leucht, S., Meyer-Lindenberg, A., Kahn, R., Rujescu, D., Sommer, I., Winter, I., McGuire, P., 2015. Magnetic resonance imaging and the prediction of outcome in first-episode schizophrenia: a review of current evidence and directions for future research. Schizophr. Bull. 41 (3), 574–583.

DeLisi, L.E., Fleischhacker, W.W., 2016. How precise is precision medicine for schizophrenia? Curr. Opin. Psychiatry 29 (3), 187–189.

Deng, C., Dean, B., 2013. Mapping the pathophysiology of schizophrenia: interactions between multiple cellular pathways. Front. Cell. Neurosci. 7, 238.

Derks, E.M., Allardyce, J., Boks, M.P., Vermunt, J.K., Hijman, R., Ophoff, R.A., GROUP, 2012. Kraepelin was right: a latent class analysis of symptom dimensions in patients and controls. Schizophr. Bull. 38 (3), 495–505.

Dickinson, D., Pratt, D.N., Giangrande, E.J., Grunnagle, M., Orel, J., Weinberger, D.R., Callicott, J.H., Berman, K.F., 2017. Attacking heterogeneity in schizophrenia by deriving clinical subgroups from widely available symptom data. Schizophr. Bull. 2017 Mar 20. https://doi.org/10.1093/schbul/sbx039.

Dluhoš, P., Schwarz, D., Cahn, W., van Haren, N., Kahn, R., Španiel, F., Horáček, J., Kašpárek, T., Schnack, H., 2017. Multi-center machine learning in imaging psychiatry: a meta-model approach. NeuroImage 155, 10–24.

Dwyer, D., Cabral, C., Kambeitz-Ilankovic, L., Kambeitz, J., Calhoun, V., Falkai, P., Pantelis, C., Meisenzahl, E., Koutsouleris, N., 2017. Brain Subtyping Enhances the Neuroanatomical Discrimination of Schizophrenia. Human Brain Mapping 2017, Vancouver (June 27 (conference poster abstract)).

First, M., Botteron, K., Carter, C., Castellanos, F.X., Dickstein, D.P., Drevets, W., Kim, K.L., Pescosolido, M.F., Rausch, S., Seymour, K.E., Sheline, Y., Zubieta, J.K., 2012. Consensus Report of the APA Work Group on Neuroimaging Markers of Psychiatric Disorders. American Psychiatric Association.

Gould, I.C., Shepherd, A.M., Laurens, K.R., Cairns, M.J., Carr, V.J., Green, M.J., 2014. Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: a support vector machine learning approach. Neuroimage Clin. 6, 229–236.

Haijma, S.V., Van Haren, N., Cahn, W., Koolschijn, P.C., Hulshoff Pol, H.E., Kahn, R.S., 2013. Brain volumes in schizophrenia: a meta-analysis in over 18,000 subjects. Schizophr. Bull. 39 (5), 1129–1138.

Howes, O.D., Montgomery, A.J., Asselin, M.C., Murray, R.M., Grasby, P.M., McGuire, P.K., 2007. Molecular imaging studies of the striatal dopami-nergic system in psychosis and predictions for the prodromal phase of psychosis. Br. J. Psychiatry Suppl. 51, s13–s18.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., 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 (7), 748–751.

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., Koutsouleris, N., 2015. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. Neuropsychopharmacology 40 (7), 1742–1751.

Keshavan, M.S., Lawler, A.N., Nasrallah, H.A., Tandon, R., 2017. New drug developments in psychosis: challenges, opportunities and strategies. Prog. Neurobiol. 152, 3–20.

Koutsouleris, N., Kahn, R.S., Chekroud, A.M., Leucht, S., Falkai, P., Wobrock, T., Derks, E.M., Fleischhacker, W.W., Hasan, A., 2016. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry 3 (10), 935–946 Oct.

Kruggel, F., Turner, J., Muftuler, L.T., Alzheimer's Disease Neuroimaging Initiative, 2010. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. NeuroImage 49 (3), 2123–2133.

Liang, S.G., Greenwood, T.A., 2015. The impact of clinical heterogeneity in schizophrenia on genomic analyses. Schizophr. Res. 161 (2–3), 490–495.

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16 (6), 321–332.

Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F., 2016. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biol. Psychiatry 80 (7), 552–561.

Marquand, A.F., Wolfers, T., Mennes, M., Buitelaar, J., Beckmann, C.F., 2016. Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1 (5), 433–447 Sep.

Mourao-Miranda, J., Reinders, A.A., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., Morgan, K.D., Fearon, P., Jones, P.B., Doody, G.A., Murray, R.M., Kapur, S., Dazzan, P., 2012. Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. Psychol. Med. 42 (5), 1037–1047.

NeuroMiner 2017. https://www.pronia.eu/neurominer/. Last accessed Oct. 2, 2017.

Nieuwenhuis, M., Schnack, H.G., van Haren, N.E., Lappin, J., Morgan, C., Reinders, A.A., Gutierrez-Tordesillas, D., Roiz-Santiañez, R., Schaufelberger, M.S., Rosa, P.G., Zanetti, M.V., Busatto, G.F., Crespo-Facorro, B., McGorry, P.D., Velakoulis, D., Pantelis, C., Wood, S.J., Kahn, R.S., Mourao-Miranda, J., Dazzan, P., 2017. Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. NeuroImage 145 (Pt B), 246–253.

Nieuwenhuis, M., van Haren, N.E., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., Schnack, H.G., 2012. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. NeuroImage 61 (3), 606–612.

Orban, P., Dansereau, C., Desbois, L., Mongeau-Pérusse, V., Giguère, C.É., Nguyen, H., Mendrek, A., Stip, E., Bellec, P., 2017. Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. Schizophr. Res. (2017 Jun 7. pii: S0920-9964(17)30302-X).

Payan, A., Montana, G., 2015. Predicting Alzheimer's Disease: A Neuroimaging Study With 3D Convolutional Neural Networks (arXiv preprint arXiv:1502.02506. 2015 Feb 9).

Regier, D.A., Narrow, W.E., Clarke, D.E., Kraemer, H.C., Kuramoto, S.J., Kuhl, E.A., Kupfer, D.J., 2013. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am. J. Psychiatry 170, 59–70.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. 20, 53–65.

Schnack, H.G., Kahn, R.S., 2016. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front. Psych. 7, 50.

Schnack, H.G., Nieuwenhuis, M., van Haren, N.E., Abramovic, L., Scheewe, T.W., Brouwer, R.M., Hulshoff Pol, H.E., Kahn, R.S., 2014. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. NeuroImage 84, 299–306.

So, H.C., Sham, P.C., 2017. Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits. Bioinformatics 33 (6), 886–892.

Squarcina, L., Castellani, U., Bellani, M., Perlini, C., Lasalvia, A., Dusi, N., Bonetto, C., Cristofalo, D., Tosato, S., Rambaldelli, G., Alessandrini, F., Zoccatelli, G., Pozzi-Mucelli, R., Lamonaca, D., Ceccato, E., Pileggi, F., Mazzi, F., Santonastaso, P., Ruggeri, M., Brambilla, P., GET UP Group, 2017. Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques. NeuroImage 145 (Pt B), 238–245.

Takahashi, T., Takayanagi, Y., Nishikawa, Y., Nakamura, M., Komori, Y., Furuichi, A., Kido, M., Sasabayashi, D., Noguchi, K., Suzuki, M., 2017. Brain neurodevelopmental markers related to the deficit subtype of schizophrenia. Psychiatry Res. 266, 10–18.

van Erp, T.G., Hibar, D.P., Rasmussen, J.M., Glahn, D.C., Pearlson, G.D., Andreassen, O.A., Agartz, I., Westlye, L.T., Haukvik, U.K., Dale, A.M., Melle, I., Hartberg, C.B., Gruber, O., Kraemer, B., Zilles, D., Donohoe, G., Kelly, S., McDonald, C., Morris, D.W., Cannon, D. M., Corvin, A., Machielsen, M.W., Koenders, L., de Haan, L., Veltman, D.J., Satterthwaite, T.D., Wolf, D.H., Gur, R.C., Gur, R.E., Potkin, S.G., Mathalon, D.H., Mueller, B.A., Preda, A., Macciardi, F., Ehrlich, S., Walton, E., Hass, J., Calhoun, V.D., Bockholt, H.J., Sponheim, S.R., Shoemaker, J.M., van Haren, N.E., Hulshoff Pol, H.E., Ophoff, R.A., Kahn, R.S., Roiz-Santiañez, R., Crespo-Facorro, B., Wang, L., Alpert, K.I., Jönsson, E.G., Dimitrova, R., Bois, C., Whalley, H.C., McIntosh, A.M., Lawrie, S.M., Hashimoto, R., Thompson, P.M., Turner, J.A., 2016. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. Mol. Psychiatry 21 (4), 547–553.

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 (Pt B), 346–364.