CHAPTER

17

Bias, noise, and interpretability in machine learning: from measurements to features

Hugo Schnack

Department of Psychiatry, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands

17.1 Introduction

In studies that use machine learning methods to investigate brain disorders, a significant amount of attention is normally paid to the selection of feature preparation techniques, machine learning algorithms, and learning designs; yet the choice of data sources and how to process them is at least as important. Once a researcher has defined the target, or output variable, they have to select a set of relevant predictors, or input variables, that are expected to be associated with the target. For instance, if the aim of the study is to diagnose schizophrenia (output variable = schizophrenia: yes/no), features could be derived using different measurement instruments, such as interviews, clinical observations, neuropsychological tests, blood samples, DNA, neuroimaging, etc. Within each measurement instrument, different acquisition modalities can be chosen (e.g., within neuroimaging: magnetic resonance imaging [MRI], electroencephalography [EEG], positron emission tomography) and acquisition protocols (e.g., within MRI: structural MRI, functional MRI, diffusion weighted imaging). Finally, the raw data can be processed and analyzed in various ways. Structural MRI data, for example, can be analyzed using global quantities such as whole-brain volume or localized measures such as regional gray matter volume or regional cortical thickness. Likewise, genetic data can be analyzed at the level of individual single-nucleotide polymorphisms (SNPs) or using polygenic risk score, an aggregated quantity (Dudbridge, 2013). All these choices influence the

Machine Learning
https://doi.org/10.1016/B978-0-12-815739-8.00017-1

quality and output of our machine learning models. These issues are not unique to machine learning. However, compared to traditional statistical approaches, they have a particularly strong impact in machine learning analysis because of the following two reasons:

(1) *For the same set of features, group-level differences do not imply groups and can be separated at the individual level.* While traditional statistical analyses aim to make statements at group level (e.g., "The patient group differs from the control group … "), machine learning models usually aim to make predictions at the level of individual subjects (e.g., "This individual belongs to the patient group rather than the control group … "). In the fields of psychiatry and neurology, group effects are, because of large biological variation, often very small. Such small differences between groups can be statistically significant (in terms of *p*-value) if the sample size allows for sufficient statistical power; however, statistical significance at group level does not necessarily imply the two groups can be separated at the level of the individual (in terms of percentage of correct classifications) (Arbabshirani, Plis, Sui, & Calhoun, 2017; Schnack, 2017; Schnack & Kahn, 2016). Therefore, the choice of features and data processing approaches should take into account that, while traditional statistical analyses focus on *p*-values, machine learning analyses focus on accuracies.

(2) *Dimensionality heavily affects performance of machine learning models.* Machine learning analyses typically model input–output relationships based on many input variables, thereby increasing the risk of detecting spurious relationships ("overfitting"; see Chapter 2). To minimize this risk, machine learning analyses often include a dimensionality reduction step; this is another important difference with traditional statistics in which there is no such step. In Section 17.2 of this chapter, we investigate how different ways of dealing with this issue influences the separability of individuals from two groups.

The cascade of steps between input and output in machine learning analyses is shown in Fig. 17.1. Some of these individual steps have been discussed in the existing literature (Tandon & Tandon, 2018); it should be noted that it is the combined effect of all these steps that determines the strength of the relationship between input and output and, more generally, the reliability of the results. In this chapter, we first provide an overview of how the decisions made at each step in the machine learning pipeline can influence the quality and output of the model. Here, we devote special attention to the initial stages of the machine learning process: the choice of target and predictor variables/features, an often overlooked source of bias and noise. Next, we focus this discussion on

how different approaches to data processing can influence the model qualitatively. Finally, we illustrate some of these issues using relevant examples from the literature.

17.2 Main sources of bias and noise in machine learning

17.2.1 Choice of target variable

Perhaps the first and most important question is what should be the task to be performed. Here, we are already confronted with potential bias and noise. For instance, if we wish to diagnose schizophrenia, we implicitly assume the existence of the phenomenon of schizophrenia. Even if such phenomenon exists (Guloksuz & van Os, 2017; Kahn, 2017), it could be a very complex, high-dimensional "concept" that is not directly observable. In this case, we have to find a proxy for it, an observable phenotype close to the phenomenon of interest. This approximation can be viewed as a projection of the true "concept" in the high-dimensional space onto a lower-dimensional subspace. An example would be defining a mental disorder via the Diagnostic and Statistical Manual of Mental Disorders (DSM) or International Statistical Classification of Diseases and Related Health Problems (ICD) criteria. Here, a possible complication is that there are further ways of labeling someone with a specific disorder (e.g., a psychiatrist's diagnosis from a consultation that does not rely on the DSM or ICD). The result of this approximation is the ability to measure our phenomenon of interest (i.e., schizophrenia) at the cost of having introduced a bias: the measurable phenotype lies at some distance (bias) from what we would actually like to predict. We want to know

$$f\colon\ X \rightarrow Y, \tag{17.1}$$

but our target becomes

$$Y^{*} = Y + \text{bias} \tag{17.2}$$

Moreover, whatever method of diagnostic labeling is used, all of them involve some kind of measurement, which is inherently susceptible to noise. For instance, even when the diagnosis is made based on the same diagnostic system (e.g., DSM), different clinicians may come to different diagnostic decisions due to less than optimal interrater reliability (Regier et al., 2013). So, our target is now to predict

$$Y^{*} = Y + \text{bias} + \text{noise} \tag{17.3}$$

Therefore, as supervised learning models are trained using data labeled by expert clinicians, the "gold standard," it is unlikely that our

algorithm will be able to capture the relationship between a set of input variables (predictors) and an often biased and noisy phenotype with 100% accuracy (Schnack & Kahn, 2016). Near perfect classifiers, in fact, should raise suspicion.

In prognostic modeling, it is possible to replace diagnosis with more concrete targets (labels), such as changes in symptom scores or real-world measures of functioning (e.g., having a job or not). However, the link between these more concrete targets and the biological measures that are often used as input variables (e.g., gray matter volume derived from neuroanatomical images) may be weaker. Furthermore, predicting future outcomes is intrinsically more challenging than predicting a current state because of the—largely unknown—impact of environmental factors including life events and therapeutic interventions on the course of a disease. On the other hand, predicting continuous targets, such as changes in symptom scores, may be more reliable than predicting dichotomous targets (e.g., presence or absence of clinical remission) because it avoids the often artificial division between categories (Janssen, Mourão-Miranda, & Schnack, 2018).

17.2.2 Choice of features

A comparable story applies to the input. Here, one needs to select and measure a variable that is expected to be predictive of the target; such selection and measurement, however, are inherently susceptible to bias and noise. Furthermore, the measurable variables available to the researcher or clinician are often only indirectly linked to the target. For example, variation in dermatoglyphic patterns (naturally occurring unique patterns on the epidermal ridges of hands and feet) is known to be statistically associated with schizophrenia (Van Oel et al., 2000); however, this must be an indirect (thus weak) association caused by the fact that the same genetic variation influences both dermatoglyphic patterns and susceptibility to schizophrenia. Thus, if dermatoglyphic pattern was found to be predictive of schizophrenia at the individual level, it would be wrong to conclude that the former has a causal effect on the latter. Of course, this might be true for many potential associations (in standard statistical analyses) and predictors (in machine learning analyses). Both bias and noise in the input are further increased by the use of "suboptimal" measurements, often for the purpose of time or cost reduction. Examples might include using only a subset of a cognitive test battery to estimate IQ or using a fast acquisition sequence to collect neuroimaging data. Thus, instead of our ideal prediction Eq. (17.1), we end up having to solve the following problem:

$$f^*: X^* \rightarrow Y^*, \tag{17.4}$$

where

$$X^* = X + \text{bias} + \text{noise}$$

and

$$Y^* = Y + \text{bias} + \text{noise}$$

17.2.3 Data processing

The input–output relationship may be diluted further by the often numerous processing steps to convert raw data into useable features (Fig. 17.1). From Section 17.3 in this chapter, we focus on how different measurements and data processing approaches are also susceptible to bias and noise.

17.2.4 Choice of machine learning algorithm

Finally, one needs to select a machine learning approach. The different algorithms, including their multiple sources of bias, are extensively discussed in Part II of the present book.

From the above observations, it is clear that "perfect" classifiers, capable of achieving accuracies close to 100%, are often unrealistic in the investigation of brain disorders. The aim of the rest of this chapter is to make the reader aware of the factors that influence the performance of our machine learning–based models and, more specifically, how data source and processing steps influence the results.

17.3 Data processing

Like most statistical models, machine learning modeling is characterized by a number of trade-offs. The *kind* or *nature* of the input features is one of them. Extensive processing of features may either increase or decrease bias and noise, as well as the interpretability of the findings. Choices regarding how to best create the optimal input features are not only based on statistical arguments but also depend on the purpose of the models: scientific discovery versus clinical practice. In the case of clinical practice, the aim is to make accurate predictions rather than understanding the underlying mechanisms. If a model reached 100% accuracy of prediction, therefore, interpretability would be of secondary importance. However, there are virtually no models in neurology and psychiatry that reach such accuracy at present. Most models have modest performance, which is too low for translation in real-world

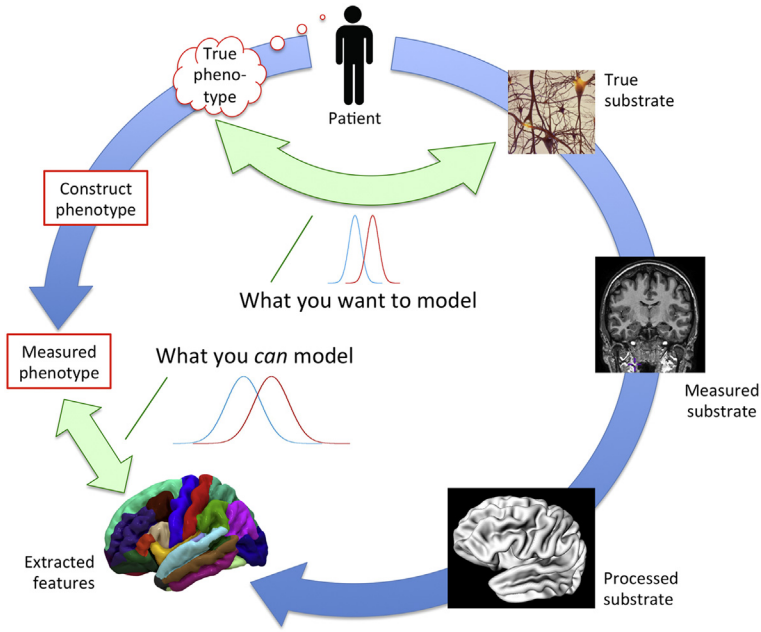

FIGURE 17.1 The weakest link—or the long route between predictors and phenotype. The goal of a machine learning approach is to make predictions about a patient, based on the (hypothesized) strong association between a patient's true phenotype (e.g., having the disorder schizophrenia) and some true substrate (e.g., some neural pathophysiology). This is depicted by the *green arrow (light green arrow in print version)* ("What we want to model"). However, usually we do neither have the true phenotype nor the true substrate at hand. We have to work with constructed phenotypes (schizophrenia is a concept) which we have to assess by doing measurements (e.g., interviews). The substrate has to be assessed as well by measurement (e.g., an MRI brain scan), which is usually processed after which features can be extracted. Instead of the ideal relationship, which would allow excellent separability (diagnosis of schizophrenia), we have to work with a highly diluted, biased, noisy, relationship—depicted by the *green arrow (light green arrow in print version)* ("What we *can* model"), providing much worse separability (largely overlapping distributions → rather inaccurate diagnosis).

clinical practice. In contrast, in the case of scientific discovery, the main endeavor is to understand the disease mechanisms rather than reaching perfect accuracy. Here, the interpretability of models is thus of critical importance.

Perhaps the main trade-off relates to the *number* features being used; this number can be somewhere between very low—corresponding to higher bias, lower risk of overfitting, thus more robust results—and very high—corresponding to lower bias, higher risk of overfitting, thus less robust results (see Chapter 2). In brain disorders research, most measurement instruments produce high- to ultrahigh-dimensional data. Examples include the Positive and Negative Syndrome Scale (PANSS)

with 30 items, neuroimages comprising tens of thousands of voxels, and genetics with 10^6 or more SNPs. There are several ways in which these data can be inputted into a machine learning model; these range from entering high-dimensional raw or preprocessed data to working with a handful of greatly reduced features. In this section, we first provide a useful way to quantify the separability of individuals from two groups, followed by a discussion on some of the different ways to address high dimensionality and their implications with respect to separability of the two groups, noise, bias, and interpretability.

17.3.1 Measuring separability between two groups

The precondition for a successful classification task, i.e., separating individuals from two (or more) groups, is that there must be a difference between the individuals from the groups with respect to some variable v. When performing this task, we only have access to the measurement of this variable, which will be our (raw) feature x. If we have two groups, g_A and g_B, we can define their difference with respect to this feature as

$$\Delta x = x_B - x_A \tag{17.5}$$

and the corresponding effect size as

$$d = \Delta x / SD \tag{17.6}$$

where SD is the pooled standard deviation,

$$SD = \sqrt{S_A^2 + S_B^2} \tag{17.7}$$

and S_A^2 and S_B^2 are the within-sample variances of x for groups A and B, respectively. The larger the effect size d, the better the separation. If the x values are normally distributed (i.e., S_A^2 and S_B^2 are the variances of a normal distribution), then separability, in terms of nonoverlap of the distributions of x for groups A and B (Cohen, 1988), can be defined as

$$\text{Accuracy} = \Phi(d/2), \tag{17.8}$$

where $\Phi(.)$ is the cumulative normal distribution function. The accuracy ranges between 0 (fully overlapping distributions) and 1 (completely nonoverlapping distributions). It follows that larger Δx and/or smaller SD will, via d, lead to better separation (Schnack, 2017; Schnack & Kahn, 2016). Note that, in traditional group-level analyses, increasing N enhances the statistical power of the study, thereby lowering the p-value; in this case, an effect with small effect size can become statistically significant; however, this does not affect separability at all because the (non)overlap of the two distributions is independent of N (Schnack, 2017).

In the investigation of brain disorders, we usually have rather small effect sizes. Based on our current knowledge of psychiatric and neurological disorders, there is no reason to assume that we could make useful predictions based on a single variable. That is why we use many, say M, variables (for which we expect nonzero d) and let a machine learning algorithm find the pattern, i.e., the weighted combination of variables that can separate the groups. In case of linear models, the M variables x_j are added to increase the effect size:

$$y = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_Mx_M = \sum_{j=1}^{M} w_jx_j \tag{17.9}$$

with a resulting effect size

$$d_{ML} = \Delta y / SD_y \tag{17.10}$$

where $\Delta y = y_B - y_A$ and SD_y is the pooled standard deviation. To see that we may expect an increased effect size d_{ML} as compared to the individual effect sizes d_j, we have to examine the effects of adding features on Δy and SD_y separately. The feature differences (Δx_j) add up linearly in Δy, and, in the case of comparable variables and weights in Eq. (17.9), $\Delta y \propto M \times \Delta x$. For SD_y, the story is different: Random noise (SD_j) in each x_j does not add up linearly (positive and negative random terms cancel out partly), but much slower: adding independent, same-sized noise leads to $SD_y \propto \sqrt{M} \times SD$. Thus, $d_{ML} \propto M/\sqrt{M} \times d$ and one gains a factor $\sqrt{M}$ compared to the single-variable effect size d by adding M independent measures.

17.3.2 Addressing high dimensionality

The virtue of machine learning methods lies in the fact that they can find the optimal combination of features to obtain the largest possible effect size d_{ML} by throwing away part of the variables and giving others weights in accordance with their effect size. That is, variables x_j with larger effect size d_j obtain larger weights. Ideally, this would result in the optimal classification model, with the highest performance. However, if we have many noisy variables x and a comparatively small number of subjects, i.e., if $M \gg N$, the modeling is prone to overfitting and poor generalization (see Chapter 2). This is why a number of automated feature reduction methods have been developed. An often used approach is feature selection, where each initial feature can be either kept or eliminated depending on whether it is deemed important to the task or not (see Chapter 2 and 6). Alternatively, it is also possible to create a new set of features from the raw data with lower dimensionality. This can be achieved by aggregating features based on previous knowledge or using data-driven unsupervised techniques such as principal component

analysis (PCA) (see Chapter 12). Both feature selection and unsupervised methods for dimensionality reduction have been covered elsewhere in this book. In this section we take a closer look at feature aggregation.

17.3.2.1 *Feature aggregation*

Combining a selection of features into one new aggregated feature has the advantage of reducing noise, while lowering the number of features to be used in the machine learning procedure. An example would be aggregating items from a clinical questionnaire into themes or factors, or MRI voxels into predefined anatomical regions, and use these as the new set of features. Contrary to automated techniques, however, this approach assumes knowledge about the relationship between features and target, i.e., we drop item- or voxel-level data and assume that a global score for "positive psychotic symptoms" or "volume of the right hippocampus," for example, is related to the output. This seems contradictory as the reason we use machine learning in the first place is precisely because we do not know the relationship between the predictor variables and output. Thus, when creating aggregates, while we are most likely improving on the model fitting by decreasing the level of noise (SD) in the data, we are also introducing bias to the model which may result in decreasing the difference between the groups (Δx): Δx_j's may cancel out, i.e., adding items together, for example, may cancel out differences that existed between groups at item level. From Eq. (17.6) or (17.10), we have to hope that reduction of noise is larger than the increase of the bias (Δx_j's canceling out) (see Eq. 17.4), so that the net effect is an increased d_{ML}. To illustrate these points, let us look at a simple numerical example of feature aggregating.

Suppose we have $M = 10$ independent "raw" features x_j that we hope to be related to, i.e., predictive for, an output y, e.g., being ill or not. For simplicity, let us assume the noise to be the same for all features: $SD_j = SD$. This feature set is schematically shown in Fig. 17.2A.

We consider the following three cases.

Case 1. All features have the same group difference Δx and thus effect size $\Delta x/SD$. Ideally, i.e., with enough subjects in our training sample (large N), the machine will find the solution:

$$y = \sum_{j=1}^{M} w_j x_j \tag{17.11}$$

with equal weights w_j. The separability, expressed as effect size (see Eq. 17.8), is then

$$d_{ML} = \sqrt{M} \times d_0 = \sqrt{10} \times d_0, \tag{17.12}$$

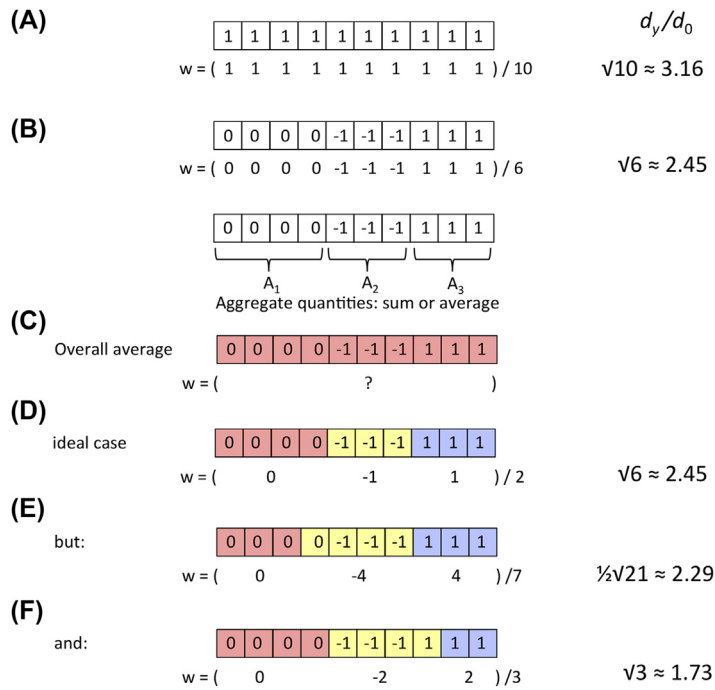

FIGURE 17.2 Example of a set of 10 features to be used to classify cases into two different groups (see Section 17.3). Groups are labeled 0 and 1. All features are assumed to be independent and have equal standard deviation; the effect sizes $d_j = \Delta x_j / SDj$ are given in terms of some arbitrary d_0. (A) The effect size for discrimination between the two groups is the same ($d_j = 1$) for all 10 features. Weights (w) will be the same (1/10) as well. Weights for this case and the following cases (B–F) are determined by maximizing the resulting effect size d_y; the scaling of w is determined by requiring $\Delta y = 1$ (i.e., 1–0). (B) Different features have different effect sizes (−1, 0, or +1). The ideal weight vector has corresponding weights. (C–F) Aggregate features (A_i) are made by summing or averaging individual features; colors indicate which features are taken together. (C) All 10 features are averaged: the features cancel out and there is no net effect in the aggregate feature. The machine learning algorithm cannot find a valid weight vector to separate the groups. (D) Ideally, the aggregation combines features having the same relationship with the target (i.e., the correct labels); the model is comparable to that of (B) (i.e., effect size d_y is the same), but the algorithm can work with a feature vector with lower dimensionality and less noise. (E) A slightly different aggregation resulting in a mild form of effect dilution in the yellow aggregate. The effect size d_y, and thus separability, of the model is lower as compared to that of (D). (F) This aggregation suffers from a somewhat more severe form of dilution in the yellow (light gray in print version) aggregate: two features' effects cancel out completely, resulting in an even lower effect size d_y and thus poorer separability.

with $d_0 = \Delta x / SD$.

If we would have aggregated the 10 features ourselves

$$x_a = \sum_{j=1}^{M} x_j \tag{17.13}$$

and then have the machine run on this single feature (which would be a very simple job: $y = x_a$), we would have obtained the same result:

$$d_{ML} = \Delta x_a / SD_{xa} = M\Delta x_a / (\sqrt{M} \times SD)$$

For small (N) datasets, this could lead to a more robust model.

Case 2. More realistically, some features (say, m_1 of them) may have a positive association with y, some (m_{-1}) have a negative association, and the remaining features (m_0) are not associated with y. Fig. 17.2B shows an example for $m_0 = 4$, $m_{-1} = m_1 = 3$. If the machine can figure out the optimal weights, ($w_{1-4} = 0$, $w_{5-7} = -1$, $w_{8-10} = +1$), the resulting d_{ML} would be $\sqrt{6} \times d_0$ (note that only features with an effect contribute). If, on the other hand, we would add all features, our x_a would have a d_{ML} of 0: positive and negative features cancel out before they can be informative (Fig. 17.2C). This might be an extreme example, but when MRI-derived whole-brain volumes are calculated, all voxels are simply added (see Section 17.4.1 for more advanced ways of machine learning and MRI analyses).

Case 3. Most of the time, we will be in an in-between situation, where we neither use the raw features nor add all features together (Fig. 17.2D–F). In practice, raw features will not be independent. Two "nearby" features are likely to measure, in part, the same quantity and/or are associated because the two quantities they assess are associated. The gain in effect size by combining features is thus less than what would be the case for completely independent features. More subtly, this also plays a role when interactions or heterogeneity are present. Suppose an outcome (e.g., disorder) requires the combined effects in features x_1 and x_2. For instance, x_1 is reduced by the amount c, which has to be accompanied by an increase of amount c in x_2. Formally, this can be expressed as (without noise):

$$x_1(\text{patient}) = x_1(\text{control}) - c,$$
$$x_2(\text{patient}) = x_2(\text{control}) + c,$$

Adding the two features, the simple aggregate

$$x_a = x_1 + x_2$$

will not be informative for the disorder because the effects $-c$ and $+c$ will cancel out.

Using both features as possible predictors will not throw away any information, but the aggregate

$$x_{LI} = x_1 - x_2$$

will have an effect $-2c$. At the same time, noise tends to cancel out, and thus the effect size may increase by a factor $\sqrt{2}$ as compared to that of the individual features, and with reduced dimensionality (1 instead of 2),

thus complexity. Of course, deciding to use the aggregate feature x_{LI} requires knowledge or a hypothesis about the roles of x_1 and x_2 in the disorder. Note that using x_a should also be based on such knowledge. A solution could be to use

$$\{x_a, x_{LI}\}$$

as feature set. While mathematically conserving the information in the features, it allows for detection of "parallel" and "antiparallel" effects. An example from brain imaging machine learning is the habit to add the volumes of left and right sides of regions of interest (ROIs) and use their sum as features. The sum of the left and right ROI volumes exhibits the same, usually large, natural variation between subjects, which may be large as compared to possible disease effects. The difference between the two, divided by their sum, is called the laterality index (*LI*). For some disorders, a change in some ROIs's *LI* is found (e.g., schizophrenia; Francis et al., 2012). In such cases, using *LI*-based features could be more powerful.

17.3.2.2 *Unsupervised methods*

Another solution is to have an unsupervised method to aggregate features, based on the covariance between them. This is how PCA approaches work. The advantage of such approaches is that they can make optimal combinations (weighted sums) of features, thus avoiding diluting/canceling out. The disadvantage, on the other hand, is that the aggregated features may not be related to the specific task. For instance, often the largest PCA components relate to general properties shared by all subjects in the sample, instead of being specifically related to the illness under investigation. Illness-related features might be expressed in the (very) small components or distributed over many components. In the first case, trimming the feature set by including only the larger PCA components will remove task-related features. In the second case, sensitivity may have been lost. Apart from these drawbacks, PCA reduction does have the advantages of greatly reducing the number of features. Moreover, if the largest PCA components are found to be related to the task, they may be interpretable.

17.4 Applications to brain disorders

In the previous section, we discussed how some of the different ways to aggregate data can influence bias and noise in our machine learning model in theory. We now illustrate these issues and their impact on the interpretability and generalizability of models using real examples from

the literature. We focus on three emerging themes in brain disorders research: prediction of brain age from neuroimaging data, using language as marker of illness, and the use of clinical variables for prognostic prediction. Each one relies on very different sources of input data, measurement instruments, and data processing, which in turn have their own advantages and challenges with regard to potential sources of bias and noise, interpretability, and generalizability.

17.4.1 Using neuroimaging to predict brain age

Many studies have investigated the possibility of predicting a person's age based on their brain images using structural MRI (sMRI) (Franke, Ziegler, Klöppel, & Gaser, 2010), functional MRI (fMRI) (Dosenbach et al., 2010), or event-related potentials (Ravan, Reilly, Trainor, & Khodayari-Rostamabad, 2011). The idea behind these efforts is that an increased age estimate of the brain may indicate the presence of a psychiatric disorder (Koutsouleris et al., 2014), while an accelerated aging of the brain, as revealed by longitudinal data, may be related to an unfavorable course of the disease (Schnack et al., 2016), for example. As discussed in Section 17.2, these brain age estimates are susceptible to bias. An important question in this regard would be if, within a certain data modality, the choice of features influences bias and, more generally, the reliability of the prediction. MRI scanners produce "raw" images consisting of voxels, 3D picture elements (cubes of $\sim$1 mm^3), which summarize the information about local brain morphology (sMRI) or activation (fMRI). This is a form of bias itself at microscopic level (Fig. 17.3A,B). From here, there are at least four routes that can be taken to balance the trade-off/compromise between bias and noise.

Aggregated voxels: Usually, images are first preprocessed and segmented into gray and white matter and other tissue classes (Fig. 17.3C). As shown in Eq. (17.13), in high-dimensionality problems, adding voxels to obtain volumes of larger structures (e.g., total brain, hippocampus, frontal lobe, total gray matter; Fig. 17.3D) is a viable option for dimensionality reduction. However, although this will decrease the level of noise in the data, it will also result in increased bias.

Cortical surface: Reconstructing the cortical surface from the voxels, resulting, for example, in local vertex-wise thickness of the cortex (Kim et al., 2005; Dale, Fischl, & Sereno, 1999) (Fig. 17.3E) improves interpretability and is potentially less biased when compared with using voxels: cortical surface is closer to the neurobiological reality. However, fitting a cortical model to limited local raw data may be prone to overfitting as such model would be sensitive to local variations and thus lead to increased noise. Indeed, the reliability of local vertex-wise cortical

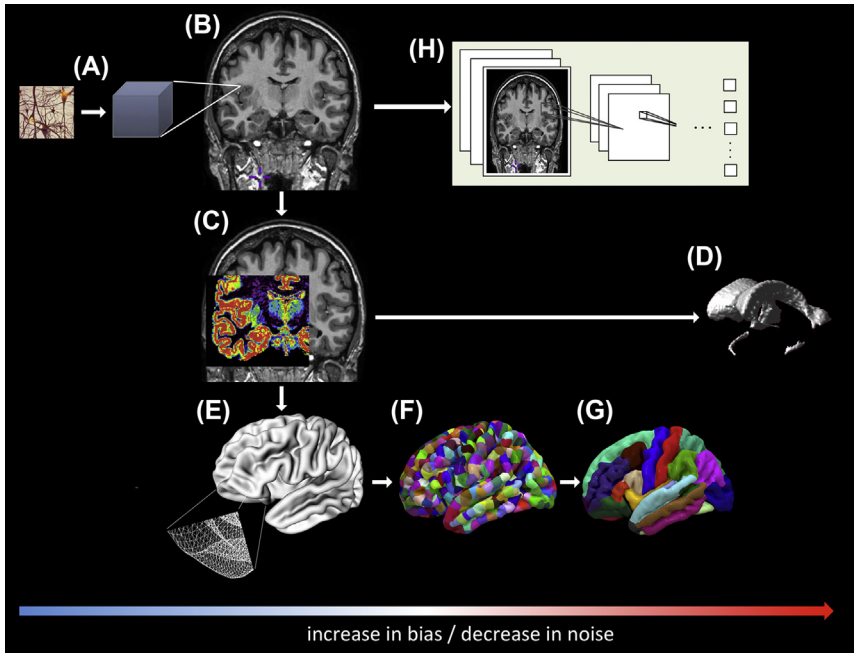

FIGURE 17.3 The use of MRI scans as data source to make predictions. Influence of processing steps and feature formation on bias, noise, interpretation. (A) An MRI scan consists of voxels, little ($\sim$1 mm^3) cubes summarizing the tissue content it represents by a single value. (B) Coronal slice of a 3D T1-weighted MRI brain scan, containing $\sim 10^6$–10^7 voxels in total (C) Processed (i.e., segmented). scan yielding high-resolution quantitative information about the brain tissue, which can be analyzed voxel-wise (e.g., using VBM). (D) Global volumes of brain structures or tissue (e.g., whole brain, ventricles (shown) or gray matter) can be calculated by directly adding the contributions from the voxels part of the structure. (E) Cortical surface reconstruction by a high-resolution triangular net, allowing for vertex-wise analysis of, e.g., cortical thickness. (F) Cortical division in (up to) 998 regions of interest (ROIs), using the Lausanne atlas (Cammoun et al., 2012). (G) Cortical division in 70 ROIs using the Desikan-Killiany atlas (Desikan et al., 2006). (H) Deep learning analysis of the raw MRI scans with convolutional neural networks (CNNs), yielding (hidden) higher-level representations of the brain image.

thickness has been shown to be lower than that of voxel-based gray matter density (VBM) (Schnack et al., 2010). Note that, if regularization is invoked to reduce the surface reconstruction's sensitivity to noise, the positive effect on bias will be reduced.

Atlas-based ROIs: Vertices from the previous step could be aggregated or averaged to obtain features such as the surface area or mean cortical thickness of certain brain regions (ROIs); this will result in a decrease in noise. However, these regions are usually automatically defined through a brain atlas (e.g., Cammoun et al., 2012; Desikan et al., 2006) (Fig. 17.3F,G)

and suffer from two issues (mentioned in Kanai, 2016): (i) unreliability (noise) due to measurement but mostly due to between-human variation in brain topology; (ii) ROIs have been defined based on a certain criteria (e.g., cell type, connectivity, or functioning), which may not be one-to-one related to the output of interest, in our current example, brain aging.

Unsupervised dimensionality reduction: It is also possible to apply an unsupervised technique, such as PCA, to all voxel or vertex values and highly reduce the dimensionality. However, a comparison between brain age models based on VBM and on VBM coupled with PCA showed no great differences in performance (Schnack et al., 2016). This is possibly due to the fact that, by performing PCA, despite reducing dimensionality and noise, there is an increase in bias. Notably, an in-between flavor is often used for voxel- and vertex-based analyses. This consists of smoothing or blurring and resampling the images or reconstructed cortical nets to lower resolution to lower the dimensionality of the feature set (Ashburner & Friston, 2000). The idea is that image elements close to each other carry more or less comparable information. Smoothing also results in noise reduction; however, it may increase bias.

Deep learning: Very recently, innovative deep learning techniques (see Chapters 9 through 11) have been applied to MRI scans with very limited preprocessing to predict age with comparable performances to that of models on processed data (Cole et al., 2017) (Fig. 17.3H). Cole and colleagues employed convolutional neural networks (CNNs; see Chapter 10), which combine image processing and prediction, thereby constructing higher-level aggregates in the hidden layers. Nevertheless, CNNs require the specification of several hyperparameters; inclusion of autoencoders (see Chapter 10) could lead to an automatic, data-driven, definition of higher-level aggregate features (potentially with lower bias) and may reduce the user input. Currently, deep neural network models are difficult to interpret.

17.4.2 Language as marker of psychiatric and developmental disorders

Higher computational power and developments in machine learning algorithms have led to increasing interest in using language as marker of psychiatric and developmental disorders (see de Boer et al., 2018 for review). Language can be used in several forms, including spoken word, written text, or markers of language knowledge. Written text has been used, for example, for author attribution based on style features (Hoorn, Frank, Kowalczyk, & Ham, 1999). The idea is that authors or, more generally, individuals each have their own language signature that can be found as a hidden pattern in text or speech. The features that determine

these signatures range from simple word counts to more complex measures related to, for example, language coherence (Elvevåg, Foltz, Weinberger, & Goldberg, 2007). Previous studies have used coherence of spoken language and other semantic and syntactic features to predict transition to psychosis in young people at risk (Bedi et al., 2015; Corcoran et al., 2018). However, the sample size was very small ($N = 34$, with 5 subjects transitioning), which, in combination with the very large number of available predictors, limited the generalizability of the prediction model. Here, lowering the number of features by using aggregate features, for example, combinations of different language features, could potentially lead to more robust models with increased generalizability.

The opposite could also be true. While studying aberrant language development in infants, one way to look at it is from the point of view of vocabulary. A frequently used instrument is the MacArthur–Bates Communicative Development Inventories (CDI; Fenson et al., 1993), which allows counting the number of words an infant knows in 22 categories. From a linguistic point of view, these 22 categories can be combined into 4 aggregates: 3 "meaningful" aggregates, covering certain word types, plus a rest category (Koster et al., 2005). Infants at family risk (FR) of developmental dyslexia (DD) and typically developing (TD) children show differences in aggregate-vocabulary sizes—on average (Koster et al., 2005). Based on this group-level knowledge, Chen, Wijnen, Koster, and Schnack (2017) set out to predict an infant's class (TD or FR) by applying a Support Vector Machine to the aggregate word counts. It turned out that the predefined aggregates did not show strong enough effects, probably because they had not been designed for detecting (risk of) DD. A subsequent machine learning analysis on the full-resolution 22 vocabulary data gave reasonable classification results of up to 68% accuracy (Chen et al., 2017). However, only some word categories from each aggregate category turned out to contribute to the classification. This suggests that aggregating features may have diluted the effects of single features—a typical example of bias. It should be noted that this study did not use a replication sample to test the generalizability of the classifier. Furthermore, the use of machine learning to differentiate between TD and FR children is of little clinical utility. In contrast, being able to predict DD before its behavioral manifestation would be of much greater clinical utility; this would require collection of data at multiple time points within a longitudinal design.

17.4.3 Using clinical variables for prognostic prediction

When making predictions about prognosis of patients with psychiatric disorders, variables obtained from clinical interviews and test batteries that assess a wide range of symptoms and functioning are likely to be

important features. For example, when acquiring data from patients with psychosis, the following data might become available: sociodemographic measures (e.g., age, sex, SES), cognitive measures, symptoms, global functioning, medication, substance use, and so on. Each of these data sources will yield one or more, up to 50+, subscores (items) or raw features, leading to a total number of features of >100. Many of these features are prone to (interrater) measurement uncertainty (noise), which can be addressed in several ways. For example, repeating a measurement will generally lower the noise; however, this costs time and money. A workaround could be the use of longitudinal data, which are often present. Although not truly measuring the same variable at the same time, it may help improving its reliability, depending on the time interval—with the potential advantage of being able to insert dynamic data in the models. An alternative could be to average scores that are assessing the same construct in different ways, for instance, clinician-rated and self-rated need of care in the CANSAS (Camberwell Assessment of Need Short Appraisal Schedule). Although these two may differ (bias), variation (noise) may be partly canceled out.

For many symptom scales, subtotal scores are calculated by adding scores on individual items. An example is the PANSS scale, for which total positive, total negative, and total general symptom scores are calculated, thus reducing the original dimensionality of 30 by a factor 10. This is great from a noise/overfitting point of view, but these sum scores have been developed for diagnostic purposes and may (likely) not optimally relate to our prediction task, e.g., predicting illness course. Solutions to this are, e.g., factor analysis or UPSM (Uncorrelated PANSS Score Matrix; Hopkins et al., 2018). The latter approach transforms the 30 PANSS scores into seven factors—representing discrete clinical domains—and has been designed to minimize between-factor correlations, which has been shown to be an especially effective approach for capturing changes in certain clinical domains (e.g., disorganization)—an example thus of feature aggregation (reduction) dedicatedly tuned to the task, in this case making prognostic predictions. Note, that for any other prediction task, new ways of aggregating might be necessary. In summary, dimensionality and noise decrease, while bias does not increase. A previous study based on raw features has used sociodemographic information, disease course, treatment adherence and response, psychiatric comorbidity, and functional and cognitive deficits to predict treatment outcome at 4 and 52 weeks (Koutsouleris et al., 2016); another study has used sociodemographic information, symptom severity, clinical, cognitive, genetic, and environmental data to predict longitudinal symptomatic and functioning outcomes at 3 and 6 years (De Nijs et al., 2018). Both studies have provided promising results, with cross-validated prediction accuracies higher than 70%.

17.5 Conclusion

The choice of data sources and the processing steps can have significant impact on the performance of a machine learning algorithm, via two main factors:

Bias: The features (predictors) and/or the target (diagnosis/prognosis) is often different from the true, optimal variables of interest. The main consequence of such differences is a weaker link between features and target and thus reduced model performance.

Noise: Almost every measurement inevitably leads to a degree of noise. Like bias, noise reduces the performance of the model.

The difference between the two causes of performance reduction is that bias reflects inherent loss of information (due to choosing the "wrong" variables or processing them in a suboptimal way), while noise could be seen as a random disturbing factor that can be addressed by acquiring more measurements (either per subject or by including more subjects). By choosing certain data sources and processing steps, one can, to a certain extent, increase or decrease the effects of bias and noise and thus influence the performance of the model. In addition, the choice of features, along with the choice of the machine learning algorithm, will affect the interpretability of the model. The need of interpretability may vary depending on its purpose (i.e., scientific discovery vs. clinical use).

17.5.1 Future directions

1. Deep learning models, which remove the requirement to preprocess the data (e.g., MRI images), show promising results. However, these models require large amount of data and more supporting evidence for their usefulness in real-world context is needed. Also, we need to develop better methods for interpreting and understanding the outputs of these models.
2. Larger datasets are needed to develop reliable models. Multicenter studies offer the opportunity to collect large samples, while at the same time introducing heterogeneity that mimics the normal clinical situation, both in terms of clinical variation between patients and the use of different people or equipment to carry out the measurements. Recently, using adversarial networks, a deep learning technique, this so-called domain adaptation problem, was successfully addressed for analysis of MRI brain scans (Kamnitsas et al., 2017).
3. Longitudinal studies, while less practical and more costly, allow the acquisition of repeated measures within the same person; the availability of multiple measures can be used to estimate and/or

reduce level of noise and to assess dynamic changes within a subject, which could inform prognostic prediction.

17.6 Key points

- Along the several stages of a machine learning analysis, decisions must be made that introduce bias and noise into the model.
- The choice of features and data processing are overlooked but important sources of noise and bias.
- High dimensionality is a common issue in studies of brain disorders; however, it can be addressed using knowledge- and data-driven approaches.
- The aggregation of features based on previous knowledge is a potential strategy to reduce data dimensionality and noise but at the expense of bias.
- The use of unsupervised methods such as PCA is another potential strategy to reduce data dimensionality and noise; however, the resulting principal components may reflect general properties shared by all subjects in the sample, instead of being specifically related to the illness under investigation.
- In neuroimaging, possible ways of minimizing noise and bias during preprocessing include the use of aggregated voxels, cortical surface measures, atlas-based ROIs, or novel deep learning methods.
- Deep learning models that require nearly unprocessed data are a particularly promising avenue for minimizing bias and noise.

Acknowledgments

We thank Wiepke Cahn, Rachel Brouwer, Jessica de Nijs, Joost Janssen, Ronald Janssen, and Jelle Schutte for useful discussions.

References

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage, 145*, 137−165. https://doi.org/10.1016/j.neuroimage.2016.02.079.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry − the methods. *NeuroImage, 11*(6 Pt 1), 805−821. https://doi.org/10.1006/nimg.2000.0582.

Bedi, G., Carrillo, F., Cecchi, G., Slezak, D., Sigman, M., Mota, N., et al. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia, 1*(1), 15030. https://doi.org/10.1038/npjschz.2015.30.

de Boer, J., Voppel, A., Begemann, M., Schnack, H., Wijnen, F., & Sommer, I. (2018). Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews, 93*, 85–92. https://doi.org/10.1016/j.neubiorev.2018.06.008.

Cammoun, L., Gigandet, X., Meskaldji, D., Thiran, J. P., Sporns, O., Do, K. Q., et al. (2012). Mapping the human connectome at multiple scales with diffusion spectrum MRI. *Journal of Neuroscience Methods, 203*(2), 386–397. https://doi.org/10.1016/j.jneumeth.2011.09.031.

Chen, A., Wijnen, F., Koster, C., & Schnack, H. (2017). Individualized early prediction of familial risk of dyslexia: A study of infant vocabulary development. *Frontiers in Psychology, 8*, 156. https://doi.org/10.3389/fpsyg.2017.00156, 2017.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, J. H., Poudel, R. P., Tsagkrasoulis, D., Caan, M. W., Steves, C., Spector, T. D., et al. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage, 163*, 115–124. https://doi.org/10.1016/j.neuroimage.2017.07.059.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., et al. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry, 17*(1), 67–75. https://doi.org/10.1002/wps.20491.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage, 9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage, 31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021.

Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., et al. (2010). Prediction of individual brain maturity using fMRI. *Science, 329*(5997), 1358–1361. https://doi.org/10.1126/science.1194144.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics, 9*(3). https://doi.org/10.1371/journal.pgen.1003348.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research, 93*(1–3), 304–316. https://doi.org/10.1016/j.schres.2007.03.001.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., et al. (1993). *The MacArthur communicative development inventories: User's guide and technical manual*. San Diego, CA: Singular Publishing Group.

Francis, A. N., Seidman, L. J., Jabbar, G. A., Mesholam-Gately, R., Thermenos, H. W., Juelich, R., et al. (2012). Alterations in brain structures underlying language function in young adults at high familial risk for schizophrenia. *Schizophrenia Research, 141*(1), 65–71. https://doi.org/10.1016/j.schres.2012.07.015.

Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage, 50*(3), 883–892. https://doi.org/10.1016/j.neuroimage.2010.01.005.

Guloksuz, S., & Os, J. V. (2017). The slow death of the concept of schizophrenia and the painful birth of the psychosis spectrum. *Psychological Medicine, 48*(02), 229–244. https://doi.org/10.1017/s0033291717001775.

Hoorn, J., Frank, S., Kowalczyk, W., & Ham, F. V. (1999). Neural network identification of poets using letter sequences. *Literary and Linguistic Computing, 14*(3), 311–338. https://doi.org/10.1093/llc/14.3.311.

Hopkins, S. C., Ogirala, A., Loebel, A., & Koblan, K. S. (2018). Transformed PANSS factors intended to reduce pseudospecificity among symptom domains and enhance understanding of symptom change in antipsychotic-treated patients with schizophrenia. *Schizophrenia Bulletin, 44*(3), 593–602. https://doi.org/10.1093/schbul/sbx101.

Janssen, R. J., Mourão-Miranda, J., & Schnack, H. G. (2018). Making individual prognoses in psychiatry using neuroimaging and machine learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(9), 798–808. https://doi.org/10.1016/j.bpsc.2018.04.004.

Kahn, R. S. (2017). Why the concept of schizophrenia is still alive and kicking. *Psychological Medicine, 48*(02), 247–248. https://doi.org/10.1017/s0033291717002069.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., et al. (2017). Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *Lecture Notes in Computer Science Information Processing in Medical Imaging*, 597–609. https://doi.org/10.1007/978-3-319-59050-9_47.

Kanai, R. (2016). Open questions in conducting confirmatory replication studies: Commentary on Boekel et al., 2015. *Cortex, 74*, 343–347. https://doi.org/10.1016/j.cortex.2015.02.020.

Kim, J. S., Singh, V., Lee, J. K., Lerch, J., Ad-Dabbagh, Y., Macdonald, D., et al. (2005). Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage, 27*(1), 210–221. https://doi.org/10.1016/j.neuroimage.2005.03.036.

Koster, C., Been, P. H., Krikhaar, E. M., Zwarts, F., Diepstra, H. D., & Leeuwen, T. H. (2005). Differences at 17 months. *Journal of Speech, Language, and Hearing Research, 48*(2), 426. https://doi.org/10.1044/1092-4388(2005/029.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., et al. (2014). Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. *Schizophrenia Bulletin, 40*(5), 1140–1153. https://doi.org/10.1093/schbul/sbt142.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., et al. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *The Lancet Psychiatry, 3*(10), 935–946. https://doi.org/10.1016/s2215-0366(16)30171-7.

de Nijs, J., van Opstal, D., Janssen, R. J., Cahn, W., Schnack, H., & GROUP Investigators. (2018). Individualized long-term outcome prediction of psychosis in an observational study: A machine learning approach. *Schizophrenia Bulletin, 44*(Suppl. l_1), S101–S102. https://doi.org/10.1093/schbul/sby015.251.

Ravan, M., Reilly, J. P., Trainor, L. J., & Khodayari-Rostamabad, A. (2011). A machine learning approach for distinguishing age of infants using auditory evoked potentials. *Clinical Neurophysiology, 122*(11), 2139–2150. https://doi.org/10.1016/j.clinph.2011.04.002.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., et al. (2013). DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry, 170*(1), 59–70. https://doi.org/10.1176/appi.ajp.2012.12070999.

Schnack, H. G. (2017). Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2017.10.023. Oct 24. pii: S0920–9964(17)30649–7.

Schnack, H. G., Haren, N. E., Brouwer, R. M., Baal, G. C., Picchioni, M., Weisbrod, M., et al. (2010). Mapping reliability in multicenter MRI: Voxel-based morphometry and cortical thickness. *Human Brain Mapping, 31*(12), 1967–1982. https://doi.org/10.1002/hbm.20991.

Schnack, H. G., Haren, N. E., Nieuwenhuis, M., Pol, H. E., Cahn, W., & Kahn, R. S. (2016). Accelerated brain aging in schizophrenia: A longitudinal pattern recognition study. *American Journal of Psychiatry, 173*(6), 607–616. https://doi.org/10.1176/appi.ajp.2015.15070922.

Schnack, H. G., & Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry, 7*. https://doi.org/10.3389/fpsyt.2016.00050.

Tandon, N., & Tandon, R. (2018). Will machine learning enable us to finally cut the gordian knot of schizophrenia. *Schizophrenia Bulletin, 44*(5), 939–941. https://doi.org/10.1093/schbul/sby101.

Van Oel, C., Baaré, W., Pol, H. H., Haag, J., Balazs, J., Dingemans, A., et al. (2001). Differentiating between low and high susceptibility to schizophrenia in twins: The significance of dermatoglyphic indices in relation to other determinants of brain development. *Schizophrenia Research, 52*(3), 181–193. https://doi.org/10.1016/s0920-9964(01)00153-0.