CHAPTER

3

Applications of machine learning to brain disorders

Cristina Scarpazza[1, 2], *Lea Baecker*[1], *Sandra Vieira*[1], *Andrea Mechelli*[1]

[1] Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom; [2] Department of General Psychology, University of Padova, Padova, Italy

3.1 Introduction

The application of machine learning in health care has been anticipated for half a century (Hinton, 2018). This vision is now becoming reality, with an increasing number of studies demonstrating the value of machine learning for assisting clinical assessment (Naylor, 2018). For instance, machine learning algorithms have been shown to outperform board-certified dermatologists in the identification of images with melanoma, basal, and squamous cell carcinoma (Esteva et al., 2017). Similarly, machine learning algorithms have been found to be capable of identifying stroke in its early stages, thanks to the detection of slight movement abnormalities (Villar, Gonzalez, Sedano, Chira, & Trejo-Gabriel-Galan, 2015). While the use of machine learning in psychiatry and neurology is far less advanced compared to other areas of medicine, there are a number of potential applications which have the potential of transforming the way psychiatric and neurological patients might be diagnosed and treated in the future. The aim of this chapter is to introduce the reader to these applications.

3.2 Why are people interested in machine learning?

Machine learning has many advantages compared to the standard statistical techniques, as introduced in Chapter 1. Critical for explaining

Machine Learning
https://doi.org/10.1016/B978-0-12-815739-8.00003-1

its wide application in the last decade is the superiority of machine learning in allowing inferences at the level of the individual. Traditional (univariate) studies are based on group-level statistics, which aim to identify differences at group level (e.g., hippocampal alteration in a group of patients Alzheimer's disease [AD] relative to a group of healthy controls). A critical limitation of group-level statistics, however, is that they do not allow inferences at the level of the individual patient. This greatly limits the translational applicability of the findings (Phillips & Swartz, 2014). In contrast, machine learning allows one to make inferences at the level of the individual patient, exponentially enhancing the possibilities of translating results into clinical practice.

The application of machine learning to brain disorders is largely based on supervised learning. Within this approach, the overwhelming majority of studies have focused on classification, where the aim is to predict the class of each observation. Studies applying this approach to brain disorders can be divided into three main categories:

(i) Prediction of illness onset: studies attempting to predict whether people at clinical risk for a certain disease will or will not become ill in the future

(ii) Diagnostic evaluation: studies attempting to develop biological or cognitive markers for detecting the presence or absence of a certain disease

(iii) Prediction of outcomes: studies attempting to predict clinical outcomes in patients with a certain illness, for example, in terms of remission versus relapse

A systematic review of the current state of the art is outside the aim of this chapter and can be found elsewhere (Arbabshirani, Plis, Sui, & Calhoun, 2017; Janssen, Mourao-Miranda, & Schnack, 2018; Jollans & Whelan, 2016; Kambeitz et al., 2015; Mateos-Perez et al., 2018; Orru, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012; Vieira, Pinaya, & Mechelli, 2017; Wolfers, Buitelaar, Beckmann, Franke, & Marquand, 2015; Woo, Chang, Lindquist, & Wager, 2017). In the following sections, we discuss these three categories of studies using examples from the existing literature.

3.2.1 Prediction of illness onset

Predicting disease onset using machine learning is a growing line of research in various fields of medicine (see, for instance, Tripoliti, Papadopoulos, Karanasiou, Naka, & Fotiadis (2017) for a review on the application of machine learning to the prediction of heart failure). In brain disorders, most of the published studies have focused on prediction of AD. The onset of this disease is typically preceded by the emergence of selective memory deficits, a condition known as mild cognitive impairment (MCI); critically, only 40%—60% of patients with MCI go on to develop AD within 5 years, while the others remain stable during this period. Studies that have used machine learning and neuroimaging data to predict conversion from MCI to AD have reported accuracies ranging from 51.44% to 80.00% (Suk, Lee, Shen, & Alzheimer's Disease Neuroimaging, 2016; Young et al., 2013). Interestingly, these accuracies can be improved by adopting a multimodality approach, in which neuroimaging data are integrated with clinical measures (Suk, Lee, Shen, & Alzheimer's Disease Neuroimaging, 2015), biological measures (Young et al., 2013), or longitudinal information (e.g., longitudinal atrophy; Lee et al., 2016).

While dozens of studies have focused on predicting transition from MCI to AD, fewer studies have attempted to predict the onset of psychiatric disorders. For instance, Koutsouleris et al. (2012) examined whether structural brain imaging could be used for predicting transition from clinical high risk for psychosis to a full-blown episode of the illness. Individuals at higher clinical risk for psychosis can be identified based on familial vulnerability and/or subthreshold clinical symptoms, in conjunction with a decline in social or cognitive functioning. Using structural magnetic resonance imaging (MRI), the authors were able to accurately predict which participants at high clinical risk did and did not develop full-blown psychosis in the following 12 months with an accuracy of 84%. Other studies, however, have reported lower accuracies. For example, a review of studies using machine learning to predict the transition from clinical high risk to psychosis has reported inconstant accuracies ranging between 52% and 84% (Janssen et al., 2018), possibly because of differences in recruitment criteria, data acquisition, and analytical methodology. Finally, a prospective 5-year follow-up study (Foland-Ross et al., 2015), which used structural MRI to predict the onset of major depression in high-risk adolescents, reported an accuracy of 70%.

Taken collectively, the above findings indicate that it is not yet possible to predict the onset of a psychiatric or neurological disease with high levels of accuracy, with the possible exception of AD. Not only are the performances not consistently high enough for clinical translation, but also most findings so far are based on small local cohorts, which do not allow for a thorough assessment of generalizability. Nevertheless, in light of the increasing number of large-scale international studies aimed at using machine learning to predict disease onset (Chung et al., 2018), this is likely to be an expanding area in the next years. Early identification of psychiatric and neurological disorders is pivotal for enabling timely and effective treatment and may lead to a reduction in transition rates. For instance, it is now widely known that the pathological process leading to AD starts 30 years before the clinical onset of the symptoms (Pike et al., 2007);

early identification of MCI patients who will develop AD would enable clinicians to intervene in the asymptomatic stage with the aim of delaying, or even preventing, the onset of the illness. Similarly, we know that subthreshold symptoms of psychosis can be detected several years before a full-blown episode of the illness; through early detection and treatment, it may be possible to reduce transition rates in this clinical population (Guloksuz & van Os, 2018).

3.2.2 Diagnostic evaluation

Diagnostic classification is the most widely and frequently used machine learning application in medical research (for instance, Esteva et al. (2017)), including brain disorders research. Diagnostic accuracy of machine learning studies tends to be higher for neurological disorders compared to psychiatric disorders. When compared to healthy controls, several studies have been able to classify patients with AD with impressive accuracies ranging from 83% to 100% (Orru et al., 2012); patients with multiple sclerosis with accuracies ranging from 86% to 96% (Mateos-Perez et al., 2018); and patients with typical and atypical parkinsonisms with accuracies higher than 80% (Mateos-Perez et al., 2018). Diagnostic accuracy for psychiatric disorders has been less stable. For instance, accuracies for mood disorders range between 59% (Serpa et al., 2014) and 90% (Mwangi, Ebmeier, Matthews, & Steele, 2012), whereas accuracies for schizophrenia range between 57% and 95%, as summarized in a recent metaanalysis (Kambeitz et al., 2015). This variability is likely to reflect the clinical heterogeneity of psychiatric disorders, in which individuals with only marginally overlapping clinical presentations may be given the same diagnosis (see Section 4.3.2). This conclusion is supported by new emerging multisite studies, characterized by high heterogeneous samples, revealing lower (e.g., 65%) but likely more reliable accuracies (Nunes et al., 2018).

Furthermore, from a clinical perspective, often the more challenging question is not whether or not a patient has a disorder, but which is the most appropriate diagnosis; this is especially the case in psychiatry, where it may take months or even years for a clinical team to reach a firm conclusion about the specific diagnosis of a patient. Machine learning is well suited to support this conclusion in the form of multiclass prediction. Unfortunately, however, this is still an underinvestigated topic with a limited number of applications in the existing literature. Studies investigating the added value of machine learning algorithms in the differential diagnosis of neurological disorders have been reviewed in detail elsewhere (Mateos-Perez et al., 2018; Orru et al., 2012). Briefly, machine learning has been applied to assist the differential diagnosis of AD and frontotemporal dementia (one study, 92% accuracy (Orru et al., 2012)) and different forms of typical and atypical parkinsonisms (two studies, accuracies ranging from not statistically significant to 97% (Mateos-Perez et al., 2018; Orru et al., 2012)). Similarly, there have been few studies investigating the value of machine learning algorithms in the differential diagnosis of psychiatric disorders. A recent review identified three studies only, two aiming at enhancing the differential diagnosis between bipolar disorder and major depression and one between bipolar disorder and schizophrenia (Kim & Na, 2018). Classification accuracies ranged from chance level to 86%, making it difficult to draw reliable conclusions. Taken collectively, these findings highlight the need for further studies using a multiclass approach; this will be particularly useful in psychiatry where clinical diagnosis can be imperfect (Regier et al., 2013).

3.2.3 Prediction of outcomes

Prediction of outcome, either in terms of disease progression or treatment response, is a key clinical challenge in psychiatry and neurology (Burki, 2016). It is therefore unsurprising that a growing number of studies are using machine learning methods to predict clinical outcomes in individual patients with depression, psychosis, bipolar disorder, attention-deficit hyperactivity disorder, and autism (Janssen et al., 2018; Jollans & Whelan, 2016; Orru et al., 2012).

Regarding disease progression, two examples on neurological disorders are particularly noteworthy. In a pioneer study on patients manifesting aphasia after stroke (Saur et al., 2010), machine learning was applied to baseline resting-state functional MRI to predict language outcome at 6 months follow-up. Results showed that it was possible to predict which patients would show good versus bad recovery with an accuracy of 76%. Another study investigated the feasibility of predicting the insurgence of dementia from the baseline brain scans of patients with Parkinson's disease. When clinical and neuroimaging measures were combined, it was possible to predict cognitive impairment at 2 years follow-up with an accuracy of 80% (Schrag, Siddiqui, Anastasiou, Weintraub, & Schott, 2017). With respect to psychiatric disorders, a number of studies have attempted to predict the course of the illness (i.e., continue/remitting) in depression (Lythe et al., 2015; Schmaal et al., 2015; Stringaris et al., 2015) and psychosis (Mourao-Miranda et al., 2012; Nieuwenhuis et al., 2017). The accuracies were consistently moderate for depression (range: 73%–78%) and inconsistent for psychosis (range: 52%–70%), highlighting the challenge of predicting illness progression in individual patients.

The prediction of treatment response is another interesting area of research. As current pharmacological and psychological treatments for psychiatric and neurological disease are only successful in a subset of patients (Kapur, Phillips, & Insel, 2012), there is much interest in trying to optimize the choice of existing treatment options by predicting their effectiveness in individual patients (Bzdok & Meyer-Lindenberg, 2018; Gabrieli, Ghosh, & Whitfield-Gabrieli, 2015). So far, however, relatively few studies have used machine learning methods to predict treatment response (see the following reviews: Janssen et al. (2018) and Jollans & Whelan (2016)). The majority of the existing studies used brain imaging data to predict response to pharmacological therapies in mood disorders. Accuracy in discriminating between response and nonresponse ranged between 70% (Gong et al., 2011) and 89% (Liu et al., 2012), indicating contrasting results. The few available studies on treatment response to antipsychotic medication in patients with first episode of psychosis report moderate accuracies; for example, Sarpal and colleagues were able to discriminate between response and nonresponse with an accuracy of 78% (Sarpal et al., 2016). Despite the paucity of the existing literature, it is hoped that this line of research will eventually lead to tailored pharmacological and psychological interventions in psychiatry and neurology.

Taken collectively, the studies published so far indicate that the application of machine learning to neurobiological data could help develop tools for prediction of disease progression or treatment response at the individual level. This would be of critical importance for clinicians as it might help to make more effective clinical decisions in the early stages of the disorder, resulting in fewer ineffective trials and higher remission rates. In addition, such tools would also have considerable implications for the economic cost of health care, for example, by reducing the need for hospitalization.

3.3 What are the main challenges in machine learning studies of psychiatric and neurological disorders?

Psychiatry and neurology are particularly challenging domains for building machine learning–based decision support systems, due to the limited understanding of the underlying causal mechanisms resulting in diagnostic and prognostic uncertainty (Shortliffe & Sepulveda, 2018). While there are multiple challenges associated with the application of machine learning in these clinical disciplines, in this chapter, we focus on three key issues: (1) absence of biomarkers; (2) reliability of diagnosis; and (3) heterogeneity.

3.3.1 Absence of biomarkers

According to the Biomarkers Definitions Working Group, a biomarker is “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” (Biomarkers Definitions Working, 2001). A biomarker is validated if it is reliable, plausible, accurate, and reproducible across clinically relevant settings (Prata, Mechelli, & Kapur, 2014). Biomarkers research aims to enhance the biological understanding of an illness, informing the development of mechanism-driven biological therapeutics. Once established, biomarkers can also be used to support the diagnostic and prognostic evaluation of individual patients in everyday clinical practice.

In most neurological disorders, the diagnosis is highly reliable, thanks to robust biomarkers that provide objective evidence. For example, AD is characterized by increased tau and decreased amyloid protein in the cerebrospinal fluid (Olsson et al., 2016). It is thus possible to confirm clinical diagnosis by performing a lumbar puncture and testing the levels of tau and amyloid proteins. Similarly, volumetric reduction of the hippocampus is included in the diagnostic criteria for AD (Dubois et al., 2007) and is used to evaluate the efficacy of pharmacological therapies (Dubois et al., 2015). In the case of psychiatric disorders, however, there are no reliable and robust biomarkers (Prata et al., 2014; The Lancet, 2016). Although some potential diagnostic and prognostic biomarkers have been proposed (Lawrie, Olabi, Hall, & McIntosh, 2011), such as larger ventricles and reduced total gray matter in schizophrenia (Heinrichs, 2004), these suffer from limited accuracy and/or generalizability in real-life clinical settings (Kapur et al., 2012). A possible explanation is that the vast majority of biological findings in psychiatry are based on prototypical patients who are rarely observed in real-world clinical practice where clinical presentation tends to be less clear-cut (Kapur et al., 2012). A further possible explanation is that the vast majority of studies used small to moderate sample sizes, which are associated with poor levels of reliability and generalizability (Cumming, 2008; Ioannidis, 2008; J. Miller, 2009). Furthermore, many of the existing studies did not use controlled designs (Prata et al., 2014).

The lack of reliable biomarkers is a major obstacle to the development of machine learning algorithms for the diagnostic and prognostic assessment of individual patients. If there are no reliable neurobiological differences between two groups of people (e.g., patients vs. controls), then even the most sophisticated machine learning algorithm will fail to discriminate between these groups at the individual level. This may explain the lower accuracies reported in machine learning studies which focused on psychiatric compared to neurologic disorders.

3.3.2 Reliability of diagnosis

Neurological and psychiatric disorders are associated with different degrees of diagnostic reliability. In the case of neurological disorders, the availability of validated biomarkers means that it is possible to carry out a diagnostic assessment with minimal uncertainty. In contrast, in the case of psychiatric disorders, the reliability of diagnostic categories has repeatedly been called into question because of the absence of validated biomarkers, the presence of clinical heterogeneity, and the high rates of comorbidity (Kapur et al., 2012). To assist clinicians in the difficult task of determining the presence or absence of a psychiatric illness, diagnostic criteria have been developed and published in the Diagnostic and Statistical Manual of Mental Disorders (DSM), which is now at its fifth edition (DSM-5) (APA, 2013), or the International Classification of Diseases (ICD), which is now at its 11th edition (WHO, 2018). The application of these diagnostic criteria, however, does not completely remove uncertainty from diagnostic evaluations of psychiatric disorders. Interrater reliability refers to the degree to which two clinicians would independently agree on the presence or absence of a disorder when the same individual is assessed using the proposed diagnostic criteria. Interrater reliability is measured using the k statistic, a coefficient ranging from 0 to 1, where 0 indicates absence of agreement and 1 indicates a complete agreement between raters. Surprisingly, the interrater reliability is relatively low for the majority of psychiatric diagnoses; for instance, 0.69 (95% confidence interval: 0.58—0.79) for autism spectrum disorder; 0.67 (0.59—0.75) for posttraumatic stress disorder; 0.56 (0.45—0.67) for bipolar I disorder; 0.56 (0.32—0.77) for binge eating disorder; 0.54 (0.43—0.66) for borderline personality disorder; 0.46 (0.34—0.59) for schizophrenia; 0.28 (0.20—0.35) for major depressive disorder; 0.21 (0.02—0.47) for antisocial personality disorder; and 0.20 (0.02—0.36) for generalized anxiety disorder (Clarke et al., 2013; Regier et al., 2013). These estimates are consistent with the low interrater agreement reported in previous studies (P. R. Miller, 2001; P. R. Miller, Dasher, Collins, Griffiths, & Brown, 2001). Fig. 3.1 shows the interrater reliability for the main psychiatric diagnoses as reported in Regier et al. (2013).

This lack of diagnostic reliability has critical implications for the training and evaluation of machine learning algorithms. The training of a diagnostic machine learning algorithm is based on the distinction between patients and controls in the training data (see Chapter 2), which in turn relies on traditional clinical assessment. If the labeling of patients and controls in the training data is biased, i.e., the gold standard is unreliable, the training algorithm will learn a biased classification. Likewise, the evaluation of a diagnostic algorithm relies on the comparison between the observed label and the predicted label; in other words, the diagnostic

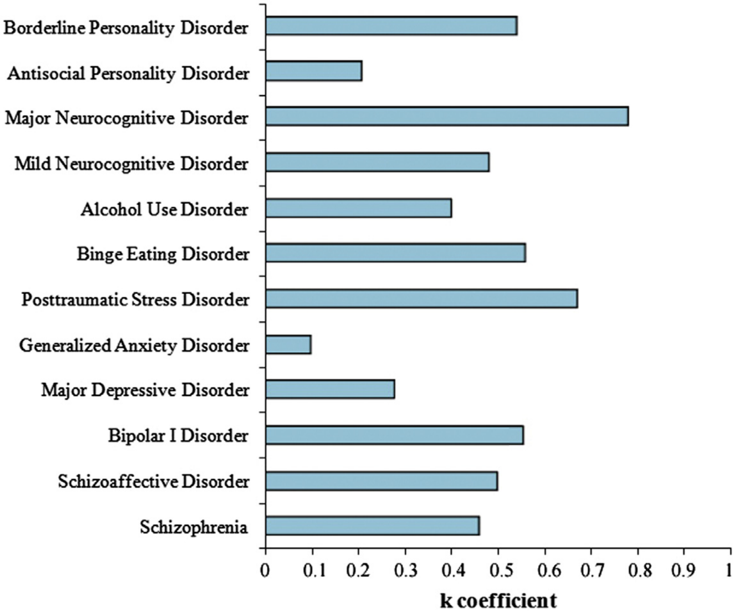

FIGURE 3.1 Interrater reliability for the main psychiatric diagnoses as reported in Regier et al. (2013).

labels based on traditional clinical assessment are compared against the diagnostic labels based on the predictions of the machine learning algorithm (see Chapter 2). However, if the original diagnostic labels are unreliable, this will lead to inaccurate estimation of the performance of the machine learning algorithm. In short, the current reliance on traditional clinical assessment for developing and evaluating diagnostic algorithms is a key challenge for the diagnostic application of machine learning in psychiatry.

3.3.3 Heterogeneity

Both psychiatric and neurological disorders tend to be heterogeneous in terms of clinical presentation, treatment response, and progression over time (Lam, Masellis, Freedman, Stuss, & Black, 2013; Logroscino, 2016; Wardenaar & de Jonge, 2013). Such clinical heterogeneity makes it challenging to use clinical data to develop machine learning algorithms for accurate diagnostic and prognostic assessment of individual patients. There is increasing evidence that clinical heterogeneity stems at least in part from distinct subtypes of patients affected by different neurobiological changes (Brugger & Howes, 2017; Lam et al., 2013; Logroscino, 2016; Wardenaar & de Jonge, 2013). For instance, a number of studies of schizophrenia have reported brain structural differences between subgroups of patients who show different clinical presentations (e.g., patients with and without hallucinations), raising the question of whether this illness should be thought of as a single disease process or a multitude of disease processes. A recent meta-analysis demonstrated that patients with first-episode schizophrenia exhibit greater intersubject variability of regional brain volume than healthy controls, particularly in the third ventricle, putamen, temporal lobe, and thalamus (Brugger & Howes, 2017). Again, such neurobiological variability makes it challenging to identify consistent patterns in the brain imaging data and use these to develop diagnostic and prognostic tools (Wolfers et al., 2015). A further complication is that both clinical and neurobiological heterogeneity are not specific to people with psychiatric and neurological disorders but are also present in disease-free individuals (Holmes & Patrick, 2018); this generates considerable overlap in statistical distributions between patients and controls, making it even harder to discriminate between the two groups at the individual level.

We should also consider the possibility that the control group in clinical studies might include false negatives, i.e., a patient affected by a psychiatric or neurological disorder was not recognized as such. To address this issue, Vernooij and colleagues investigated the presence of incidental findings (i.e., unexpected and asymptomatic brain abnormalities) in the general population by recruiting and scanning 2000 individuals who had never received a psychiatric or neurological diagnosis (Vernooij et al., 2007). Incidental findings (e.g., asymptomatic brain infarcts, arachnoid cysts, etc.) were detected in around 10% of the sample, indicating a higher than expected proportion of false negatives in the general population. This raises the possibility that machine learning studies aimed at developing diagnostic algorithms might suffer from the inclusion of undiagnosed and asymptomatic patients who have been misclassified as healthy controls.

When considering the impact of heterogeneity in machine learning studies, it is paramount to pay attention to the issue of sample size. Studies with smaller sample sizes typically have tightly controlled exclusion criteria, and therefore participants tend to be homogeneous and resemble the "ideal patient." This makes it easier for an algorithm to learn shared abnormalities in patients relative to controls, resulting in high but potentially overoptimistic accuracies. Not surprisingly, it can be difficult to replicate these accuracies in independent samples where the same tightly controlled exclusion criteria have not been applied. As sample sizes grow, so does heterogeneity because of the loosening of inclusion criteria. On the one hand, higher level of heterogeneity within a sample makes it more challenging for an algorithm to learn shared abnormalities in patients, resulting in lower accuracies (Janssen et al., 2018; Nunes et al.,

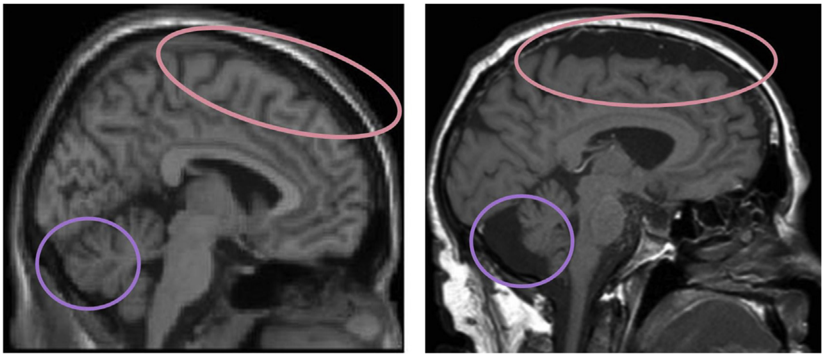

FIGURE 3.2 Neuroanatomical heterogeneity in the healthy brain. Left: Image of the brain of a 60-year-old healthy individual. The brain has no obvious abnormalities at visual inspection. Right: Image of the brain of a 54-year-old healthy individual. At visual inspection, enlargement of periencephalic spaces is evident around multiple areas including the frontal cortex and the cerebellum.

2018). On the other hand, larger samples tend to be more representative of the clinical population of interest and as such carry greater translational potential in real-world clinical practice (Nunes et al., 2018). Heterogeneity in patients and controls is, therefore, a challenging but important aspect that should be taken into account when developing machine learning algorithms to support the diagnostic and prognostic assessment of brain disorders. An example of neuroanatomical heterogeneity is shown in Fig. 3.2.

3.4 How good is good enough?

Despite the above challenges, there is increasing optimism that machine learning could be used to develop tools to support clinical diagnosis and predict prognosis and treatment response in brain disorders (Janssen et al., 2018; Jiang et al., 2017; Jollans & Whelan, 2016; Mateos-Perez et al., 2018). This raises the important question of how good is good enough. In other words, what level of performance should be achieved for a machine learning–based tool to be considered clinically useful? In medicine, the clinical utility of a diagnostic or prognostic tool mainly depends on two factors: (i) the available alternatives and (ii) the cost of misclassification.

Available alternatives: The traditional view is that, for a biomarker to be considered clinically useful, it must show an overall accuracy of 80% or more (Savitz, Rauch, & Drevets, 2013). However, this is not always feasible, for example, in the case of prognostic algorithms. This has led to the suggestion that a machine learning–based decision support system can be considered clinically useful if it enables diagnostic or prognostic inferences with an accuracy that is similar or superior to standard methods (Shortliffe & Sepulveda, 2018). In the case of AD, for example, the current clinical criteria for diagnosing "probable illness" provide a sensitivity of 85% when compared with autopsy-based cases. Thus, the sensitivity of a machine learning—based tool for diagnostic assessment of individual patents should exceed 85% to be considered clinically useful (Savitz et al., 2013). In the case of AD, the development of machine learning—based tools is facilitated by the possibility of confirming the diagnosis postmortem; this possibility, however, is not present in the case of psychiatric disorder because of the absence of robust biomarkers.

Cost of misclassification: There are no established guidelines for considering the cost of misclassification when developing machine learning—based decision support systems. Generally, a machine learning algorithm developer might select a threshold value on the basis of the algorithm's intended translational application (Perlis, 2011). In the case of diagnostic tools, the selected threshold should take into account the fact that the cost of erroneously misclassifying someone ill as healthy may be higher than the cost of misclassifying someone healthy as ill; thus, an algorithm which provides excellent sensitivity but only good specificity may be preferred to one which provides excellent specificity but only good sensitivity (Savitz et al., 2013). In a recent investigation, for example, a machine learning—based movement detecting device for early stroke recognition was designed to activate an alert if the participant's movement differed from the normal pattern (Villar et al., 2015). Here, the cost of a false negative (i.e., a patient with stroke is incorrectly classified as healthy) can lead to the death of the patient, while the cost of a false positive (i.e., a healthy individual is incorrectly classified as having a stroke) would primarily lead to performing unnecessary examinations. In other words, the clinical priority is to correctly detect the occurrence of a stroke and refer the patient for treatment as early as possible. It follows that an algorithm with higher sensitivity and lower specificity should be preferred to one showing the opposite pattern. In the case of prognostic tools, inaccurate prognosis can lead to inefficient use of clinical resources, such as the administration of treatment that is not needed and may cause adverse side effects. Therefore, even prognostic models with high accuracy may not be suitable for clinical use if the cost of misclassification is equally high. Here, a tool is useful if it provides high levels of sensitivity and the consequence of a false negative is particularly serious (Perlis, 2011); in contrast, the specificity of a tool must be prioritized if it is used to establish the need for invasive or risky interventions (Perlis, 2011). See Fig. 3.3 for an illustration of how the cost impact classification should be taken into account when developing machine learning—based decision support systems.

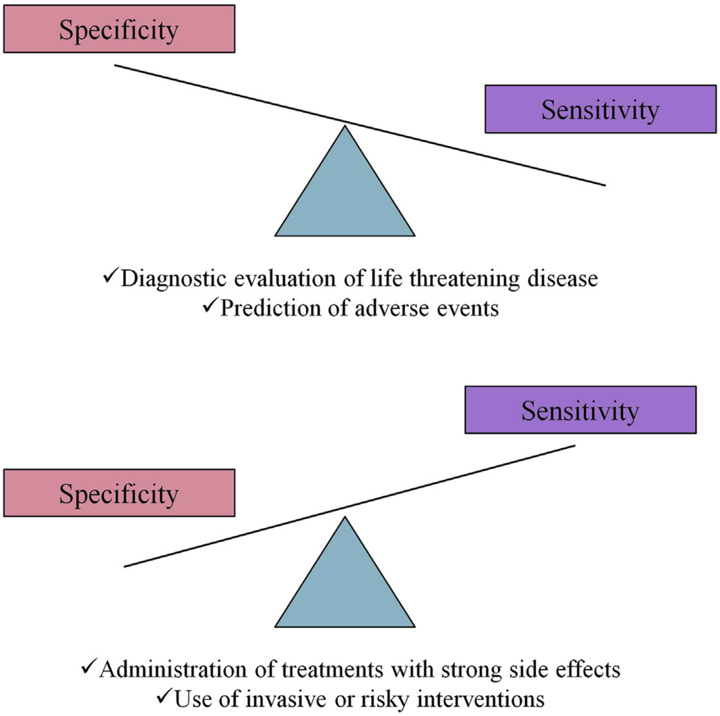
Specificity
Sensitivity
✓Diagnostic evaluation of life threatening disease
✓Prediction of adverse events
Sensitivity
Specificity
✓Administration of treatments with strong side effects
✓Use of invasive or risky interventions

FIGURE 3.3 The cost of misclassification. Upper panel: Preferred trade-off between specificity and sensitivity when the cost of a false negative is high (e.g., diagnostic evaluation of life-threatening disease). Lower panel: Preferred trade-off between specificity and sensitivity when the cost of a false positive is high (e.g., treatment has strong side effects).

Finally, even when a machine learning–based tool meets the above criteria, it is important to consider other factors within a cost/benefit framework (Savitz et al., 2013); these might include, for example, cost to clinical services, inconvenience to patients, and ethical considerations regarding the ownership and accountability of clinical decision-making (see Chapter 18).

3.5 Is machine learning ready to be applied in psychiatry and neurology?

Early promising results have suggested that machine learning techniques could be used to develop diagnostic and prognostic tools to assist clinical decision-making in psychiatry and neurology. Despite this early promise, however, the process by which a machine learning algorithm might be translated into clinical practice has received scant attention in the literature. Here, we discuss the main methodological and theoretical issues involved in the real-world implementation of machine learning—based clinical decision tools (Vieira et al. 2019).

At present, a major methodological issue is that studies have tended to use small sample sizes to train and test the machine learning algorithms; this, as discussed earlier, may lead to overoptimistic results. Furthermore, the majority of studies have used data acquired from a single cohort, which tends to be associated with poor generalizability to new cohorts. To overcome the limitations of small size and single cohort studies, an increasing number of studies are using large-scale, multisite datasets; these include, for instance, the AD neuroimaging initiative (ADNI) (Lin et al., 2018); the Parkinson's Progression Markers Initiative (PPMI) (Peng et al., 2017); and the ENIGMA Bipolar Disorder working group (ENIGMA-BP) (Nunes et al., 2018). This trend, reflecting widespread awareness of the importance of increasing the size and heterogeneity of the data to develop robust models, is a necessary step toward the future implementation of machine learning—based clinical decision tools in psychiatry and neurology.

The main theoretical issue relates to how illness is conceptualized. Firstly, we must avoid the pitfall of considering neurobiological alterations pathological per se, as we know that there are several genetic and environmental factors that can affect the structure and function of the human brain without necessarily leading to pathology (Fuchs & Flugge, 2014). Instead, researchers and clinicians must interpret the output of a machine learning model in light of the patient's clinical history and symptomatology. Second, the existence of more than one diagnostic system (DSM-5 and ICD-10) may limit the generalizability of the results obtained in studies which recruited patients based on a specific diagnostic system. Before machine learning can be applied in real-world clinical practice, we must gain a better understanding of the generalizability of the models across the different diagnostic systems. Third, the existing conceptualization of psychiatric and neurological disorders as distinct categorical entities is questioned by some who suggest that a dimensional spectrum (i.e., a continuum between health and illness and between the different disorders) may provide a better account of the clinical reality (Kapur et al., 2012). This is consistent with the emerging idea that each disorder is best understood as a combination of diagnosis-specific features and a transdiagnostic factor, reflecting general psychopathology (Caspi et al., 2014).

In light of these outstanding methodological and theoretical issues, it may come as a surprise that a number of machine learning—based tools are already available to researchers and clinicians (Azab, Carone, Ying, & Yousem, 2015; Manjon & Coupe, 2016). It should be noted, however, that all of these existing tools have been developed to support diagnostic rather than prognostic evaluation. Furthermore, all of these tools are targeted toward neurological disorders, especially AD for atrophy quantification (Brewer, Magda, Airriess, & Smith, 2009; Kovacevic, Rafii, Brewer, & Alzheimer's Disease Neuroimaging, 2009) and multiple sclerosis for automated segmentation and volume calculation of inflammatory lesions (Vrooman et al., 2007). To our knowledge, no tools have been developed and validated for use in specific psychiatric disorders, where brain alterations are more subtle and distributed.

In conclusion, to develop reliable machine learning–based tools to support clinical decision-making in psychiatry, we need larger and more heterogeneous studies as well as greater consensus on the conceptualization of the illness. Nevertheless, the current trend toward large-scale multicohort studies and the increasing recognition of the shortcomings of current conceptualizations of disease suggest that the development and validation of such machine learning–based tools for clinical use may be feasible in the coming years.

3.6 Future directions and concluding remarks

The development of machine learning–based tools to support diagnostic and prognostic evaluation of psychiatric and neurological disorders would be a critical step toward the realization of precision medicine in brain disorders. To enhance the likelihood of clinical translation, we propose the following recommendations:

(a) We should focus on developing models with high clinical utility, which in turn depends on the available alternatives and the cost of misclassification (see section 3.4);
(b) We should explore the advantage of using multimodal machine learning algorithms, which integrate clinical, genetic, social, and neurobiological data instead of focusing on a specific type of data (see Chapter 16);
(c) We should focus on symptoms rather than on diagnosis, as a potential way of overcoming the issue of diagnostic uncertainty and lack of clear boundaries between disorders (this is more relevant to psychiatry than neurology);
(d) We should focus on predicting outcomes rather than diagnosis, as the former addresses a greater clinical challenge than the latter;
(e) To maximize the generalizability of the model, we should train and test an algorithm using data collected from different cohorts.

3.7 Key points

- Machine learning techniques have three main applications in brain disorders: prediction of illness onset, diagnostic evaluation, and prediction of future outcomes
- Challenges in the application of machine learning to brain disorders include the absence of biomarkers, the poor reliability of diagnosis, and clinical/neurobiological heterogeneity
- The clinical utility of a machine learning–based tool depends on available alternatives and the cost of misclassification rather than the performance of the algorithm per se
- Machine learning is not yet ready to inform diagnostic and prognostic evaluations of psychiatric and neurological disorders in real-world clinical settings, but the increasing trend toward larger and more heterogeneous samples offers promise
- We have proposed a set of guidelines for enhancing the likelihood of clinical translation in the future

References

APA. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (fifth ed.). Arlington, VA: American Psychiatric Association.

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. https://doi.org/10.1016/j.neuroimage.2016.02.079.

Azab, M., Carone, M., Ying, S. H., & Yousem, D. M. (2015). Mesial temporal sclerosis: accuracy of NeuroQuant versus neuroradiologist. *American Journal of Neuroradiology, 36*(8), 1400–1406. https://doi.org/10.3174/ajnr.A4313.

Biomarkers Definitions Working, G. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics, 69*(3), 89–95. https://doi.org/10.1067/mcp.2001.113989.

Brewer, J. B., Magda, S., Airriess, C., & Smith, M. E. (2009). Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in Alzheimer disease. *American Journal of Neuroradiology, 30*(3), 578–580. https://doi.org/10.3174/ajnr.A1402.

Brugger, S. P., & Howes, O. D. (2017). Heterogeneity and homogeneity of regional brain structure in schizophrenia: a meta-analysis. *JAMA Psychiatry, 74*(11), 1104–1111. https://doi.org/10.1001/jamapsychiatry.2017.2663.

Burki, T. K. (2016). Predicting lung cancer prognosis using machine learning. *Lancet Oncology, 17*(10), e421. https://doi.org/10.1016/S1470-2045(16)30436-3.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., et al. (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science, 2*(2), 119–137. https://doi.org/10.1177/2167702613497473.

Chung, Y., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., et al. (2018). Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. *JAMA Psychiatry, 75*(9), 960—968. https://doi.org/10.1001/jamapsychiatry.2018.1543.

Clarke, D. E., Narrow, W. E., Regier, D. A., Kuramoto, S. J., Kupfer, D. J., Kuhl, E. A., et al. (2013). DSM-5 field trials in the United States and Canada, Part I: study design, sampling strategy, implementation, and analytic approaches. *The American Journal of Psychiatry, 170*(1), 43—58. https://doi.org/10.1176/appi.ajp.2012.12070998.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286—300. https://doi.org/10.1111/j.1745-6924.2008.00079.x.

Dubois, B., Chupin, M., Hampel, H., Lista, S., Cavedo, E., Croisile, B., et al. (2015). Donepezil decreases annual rate of hippocampal atrophy in suspected prodromal Alzheimer's disease. *Alzheimers Dementia, 11*(9), 1041—1049. https://doi.org/10.1016/j.jalz.2014.10.003.

Dubois, B., Feldman, H. H., Jacova, C., Dekosky, S. T., Barberger-Gateau, P., Cummings, J., et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurology, 6*(8), 734—746. https://doi.org/10.1016/S1474-4422(07)70178-3.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115—118. https://doi.org/10.1038/nature21056.

Foland-Ross, L. C., Sacchet, M. D., Prasad, G., Gilbert, B., Thompson, P. M., & Gotlib, I. H. (2015). Cortical thickness predicts the first onset of major depression in adolescence. *International Journal of Developmental Neuroscience, 46*, 125—131. https://doi.org/10.1016/j.ijdevneu.2015.07.007.

Fuchs, E., & Flugge, G. (2014). Adult neuroplasticity: more than 40 years of research. *Neural Plasticity, 2014*, 541870. https://doi.org/10.1155/2014/541870.

Gabrieli, J. D. E., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron, 85*(1), 11—26. https://doi.org/10.1016/j.neuron.2014.10.047.

Gong, Q., Wu, Q., Scarpazza, C., Lui, S., Jia, Z., Marquand, A., et al. (2011). Prognostic prediction of therapeutic response in depression using high-field MR imaging. *Neuroimage, 55*(4), 1497—1503. https://doi.org/10.1016/j.neuroimage.2010.11.079.

Guloksuz, S., & van Os, J. (2018). Need for evidence-based early intervention programmes: a public health perspective. *Evidence-Based Mental Health, 21*(4), 128—130. https://doi.org/10.1136/ebmental-2018-300030.

Heinrichs, R. W. (2004). Meta-analysis and the science of schizophrenia: variant evidence or evidence of variants? *Neuroscience & Biobehavioral Reviews, 28*(4), 379—394. https://doi.org/10.1016/j.neubiorev.2004.06.003.

Hinton, G. (2018). Deep learning-A technology with the potential to transform health care. *JAMA, 320*(11), 1101—1102. https://doi.org/10.1001/jama.2018.11100.

Holmes, A. J., & Patrick, L. M. (2018). The myth of optimality in clinical neuroscience. *Trends in Cognitive Sciences, 22*(3), 241—257. https://doi.org/10.1016/j.tics.2017.12.006.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640—648. https://doi.org/10.1097/EDE.0b013e31818131e7.

Janssen, R. J., Mourao-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.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology, 2*(4), 230—243. https://doi.org/10.1136/svn-2017-000101.

Jollans, L., & Whelan, R. (2016). The clinical added value of imaging: a perspective from outcome prediction. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1*(5), 423–432. https://doi.org/10.1016/j.bpsc.2016.04.005.

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., et al. (2015). Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology, 40*(7), 1742–1751. https://doi.org/10.1038/npp.2015.22.

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry, 17*(12), 1174–1179. https://doi.org/10.1038/mp.2012.105.

Kim, Y. K., & Na, K. S. (2018). Application of machine learning classification for structural brain MRI in mood disorders: critical review from a clinical perspective. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 80*(Pt B), 71–80. https://doi.org/10.1016/j.pnpbp.2017.06.024.

Koutsouleris, N., Borgwardt, S., Meisenzahl, E. M., Bottlender, R., Moller, H. J., & Riecher-Rossler, A. (2012). Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. *Schizophrenia Bulletin, 38*(6), 1234–1246. https://doi.org/10.1093/schbul/sbr145.

Kovacevic, S., Rafii, M. S., Brewer, J. B., & Alzheimer's Disease Neuroimaging, I. (2009). High-throughput, fully automated volumetry for prediction of MMSE and CDR decline in mild cognitive impairment. *Alzheimer Disease and Associated Disorders, 23*(2), 139–145. https://doi.org/10.1097/WAD.0b013e318192e745.

Lam, B., Masellis, M., Freedman, M., Stuss, D. T., & Black, S. E. (2013). Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimer's Research & Therapy, 5*(1), 1. https://doi.org/10.1186/alzrt155.

Lawrie, S. M., Olabi, B., Hall, J., & McIntosh, A. M. (2011). Do we have any solid evidence of clinical utility about the pathophysiology of schizophrenia? *World Psychiatry, 10*(1), 19–31.

Lee, S. H., Bachman, A. H., Donghyeon, Y., Lim, J., Ardekani, B. A., & Alzheimer's Disease Neuroimaging, I. (2016). Predicting progression from mild cognitive impairment to Alzheimer's disease using longitudinal callosal atrophy. *Alzheimer's Dementia: Diagnosis, Assessment & Disease Monitoring, 2*, 68–74.

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., et al. (2018). Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Frontiers in Neuroscience, 12*, 777. https://doi.org/10.3389/fnins.2018.00777.

Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z., et al. (2012). Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. *PLoS One, 7*(7), e40968. https://doi.org/10.1371/journal.pone.0040968.

Logroscino, G. (2016). Classifying change and heterogeneity in amyotrophic lateral sclerosis. *Lancet Neurology, 15*(11), 1111–1112. https://doi.org/10.1016/S1474-4422(16)30206-X.

Lythe, K. E., Moll, J., Gethin, J. A., Workman, C. I., Green, S., Lambon Ralph, M. A., et al. (2015). Self-blame-Selective hyperconnectivity between anterior temporal and subgenual cortices and prediction of recurrent depressive episodes. *JAMA Psychiatry, 72*(11), 1119–1126. https://doi.org/10.1001/jamapsychiatry.2015.1813.

Manjon, J. V., & Coupe, P. (2016). volBrain: an online MRI brain volumetry system. *Front Neuroinform, 10*, 30. https://doi.org/10.3389/fninf.2016.00030.

Mateos-Perez, J. M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., & Evans, A. C. (2018). Structural neuroimaging as clinical predictor: a review of machine learning applications. *NeuroImage: Clinical, 20*, 506–522. https://doi.org/10.1016/j.nicl.2018.08.019.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review, 16*(4), 617—640. https://doi.org/10.3758/PBR.16.4.617.

Miller, P. R. (2001). Inpatient diagnostic assessments: 2. Interrater reliability and outcomes of structured vs. unstructured interviews. *Psychiatry Research, 105*(3), 265—271.

Miller, P. R., Dasher, R., Collins, R., Griffiths, P., & Brown, F. (2001). Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Research, 105*(3), 255—264.

Mourao-Miranda, J., Reinders, A. A., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., et al. (2012). Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. *Psychological Medicine, 42*(5), 1037—1047. https://doi.org/10.1017/S0033291711002005.

Mwangi, B., Ebmeier, K. P., Matthews, K., & Steele, J. D. (2012). Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain, 135*(Pt 5), 1508—1521. https://doi.org/10.1093/brain/aws084.

Naylor, C. D. (2018). On the prospects for a (deep) learning health care system. *JAMA, 320*(11), 1099—1100. https://doi.org/10.1001/jama.2018.11103.

Nieuwenhuis, M., Schnack, H. G., van Haren, N. E., Lappin, J., Morgan, C., Reinders, A. A., et al. (2017). Multi-center MRI prediction models: predicting sex and illness course in first episode psychosis patients. *Neuroimage, 145*(Pt B), 246—253. https://doi.org/10.1016/j.neuroimage.2016.07.027.

Nunes, A., Schnack, H. G., Ching, C. R. K., Agartz, I., Akudjedu, T. N., Alda, M., et al. (2018). Using structural MRI to identify bipolar disorders - 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-018-0228-9.

Olsson, B., Lautner, R., Andreasson, U., Ohrfelt, A., Portelius, E., Bjerke, M., et al. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology, 15*(7), 673—684. https://doi.org/10.1016/S1474-4422(16)00070-3.

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews, 36*(4), 1140—1152. https://doi.org/10.1016/j.neubiorev.2012.01.004.

Peng, B., Wang, S., Zhou, Z., Liu, Y., Tong, B., Zhang, T., et al. (2017). A multilevel-ROI-features-based machine learning method for detection of morphometric biomarkers in Parkinson's disease. *Neuroscience Letters, 651*, 88—94. https://doi.org/10.1016/j.neulet.2017.04.034.

Perlis, R. H. (2011). Translating biomarkers to clinical practice. *Molecular Psychiatry, 16*(11), 1076—1087. https://doi.org/10.1038/mp.2011.63.

Phillips, M. L., & Swartz, H. A. (2014). A critical appraisal of neuroimaging studies of bipolar disorder: toward a new conceptualization of underlying neural circuitry and a road map for future research. *The American Journal of Psychiatry, 171*(8), 829—843. https://doi.org/10.1176/appi.ajp.2014.13081008.

Pike, K. E., Savage, G., Villemagne, V. L., Ng, S., Moss, S. A., Maruff, P., et al. (2007). Beta-amyloid imaging and memory in non-demented individuals: evidence for preclinical Alzheimer's disease. *Brain, 130*(Pt 11), 2837—2844. https://doi.org/10.1093/brain/awm238.

Prata, D., Mechelli, A., & Kapur, S. (2014). Clinically meaningful biomarkers for psychosis: a systematic and quantitative review. *Neuroscience & Biobehavioral Reviews., 45*, 134—141. https://doi.org/10.1016/j.neubiorev.2014.05.010.

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. *The American Journal of Psychiatry, 170*(1), 59—70. https://doi.org/10.1176/appi.ajp.2012.12070999.

Sarpal, D. K., Argyelan, M., Robinson, D. G., Szeszko, P. R., Karlsgodt, K. H., John, M., et al. (2016). Baseline striatal functional connectivity as a predictor of response to antipsychotic drug treatment. *The American Journal of Psychiatry, 173*(1), 69—77. https://doi.org/10.1176/appi.ajp.2015.14121571.

Saur, D., Ronneberger, O., Kummerer, D., Mader, I., Weiller, C., & Kloppel, S. (2010). Early functional magnetic resonance imaging activations predict language outcome after stroke. *Brain, 133*(Pt 4), 1252—1264. https://doi.org/10.1093/brain/awq021.

Savitz, J. B., Rauch, S. L., & Drevets, W. C. (2013). Clinical application of brain imaging for the diagnosis of mood disorders: the current state of play. *Molecular Psychiatry, 18*(5), 528—539. https://doi.org/10.1038/mp.2013.25.

Schmaal, L., Marquand, A. F., Rhebergen, D., van Tol, M. J., Ruhe, H. G., van der Wee, N. J., et al. (2015). Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: a multivariate pattern recognition study. *Biological Psychiatry, 78*(4), 278—286. https://doi.org/10.1016/j.biopsych.2014.11.018.

Schrag, A., Siddiqui, U. F., Anastasiou, Z., Weintraub, D., & Schott, J. M. (2017). Clinical variables and biomarkers in prediction of cognitive impairment in patients with newly diagnosed Parkinson's disease: a cohort study. *The Lancet Neurology, 16*(1), 66—75. https://doi.org/10.1016/S1474-4422(16)30328-3.

Serpa, M. H., Ou, Y., Schaufelberger, M. S., Doshi, J., Ferreira, L. K., Machado-Vieira, R., et al. (2014). Neuroanatomical classification in a population-based sample of psychotic major depression and bipolar I disorder with 1 year of diagnostic stability. *BioMed Research International, 2014*, 706157. https://doi.org/10.1155/2014/706157.

Shortliffe, E. H., & Sepulveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA, 320*(21), 2199—2200. https://doi.org/10.1001/jama.2018.17163.

Stringaris, A., Vidal-Ribas Belil, P., Artiges, E., Lemaitre, H., Gollier-Briant, F., Wolke, S., et al. (2015). The brain's response to reward anticipation and depression in adolescence: dimensionality, specificity, and longitudinal predictions in a community-based sample. *The American Journal of Psychiatry, 172*(12), 1215—1223. https://doi.org/10.1176/appi.ajp.2015.14101298.

Suk, H. I., Lee, S. W., Shen, D., & Alzheimer's Disease Neuroimaging, I. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function, 220*(2), 841—859. https://doi.org/10.1007/s00429-013-0687-3.

Suk, H. I., Lee, S. W., Shen, D., & Alzheimer's Disease Neuroimaging, I. (2016). Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Structure and Function, 221*(5), 2569—2587. https://doi.org/10.1007/s00429-015-1059-y.

The Lancet, P. (2016). Blood biomarkers in psychiatry. *Lancet Psychiatry, 3*(8), 693. https://doi.org/10.1016/S2215-0366(16)30176-6.

Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Computational and Structural Biotechnology Journal, 15*, 26—47. https://doi.org/10.1016/j.csbj.2016.11.001.

Vernooij, M. W., Ikram, M. A., Tanghe, H. L., Vincent, A. J., Hofman, A., Krestin, G. P., et al. (2007). Incidental findings on brain MRI in the general population. *The New England Journal of Medicine, 357*(18), 1821—1828. https://doi.org/10.1056/NEJMoa070972.

Vieira, S., Gong, Q. Y., Pinaya, W. H. L., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-García, V., Setien-Suero, E., Scheepers, F. E., Van Haren, N. E. M., Marques, T. R., Murray, R. M., David, A., Dazzan, P., McGuire, P., & Mechelli, A. (2019). Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence. *Schizophr Bull 2019*. https://www.ncbi.nlm.nih.gov/pubmed/30809667.

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neuroscience & Biobehavioral Reviews., 74*(Pt A), 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002.

Villar, J. R., Gonzalez, S., Sedano, J., Chira, C., & Trejo-Gabriel-Galan, J. M. (2015). Improving human activity recognition and its application in early stroke diagnosis. *International Journal of Neural Systems, 25*(4), 1450036. https://doi.org/10.1142/S0129065714500361.

Vrooman, H. A., Cocosco, C. A., van der Lijn, F., Stokking, R., Ikram, M. A., Vernooij, M. W., et al. (2007). Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. *Neuroimage, 37*(1), 71–81. https://doi.org/10.1016/j.neuroimage.2007.05.018.

Wardenaar, K. J., & de Jonge, P. (2013). Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Medicine, 11*, 201. https://doi.org/10.1186/1741-7015-11-201.

WHO. (2018). *International Classification of Diseases* (11th ed.). World Health Organization.

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews, 57*, 328–349. https://doi.org/10.1016/j.neubiorev.2015.08.001.

Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience, 20*(3), 365–377. https://doi.org/10.1038/nn.4478.

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., Ourselin, S., & Alzheimer's Disease Neuroimaging, I. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical, 2*, 735–745. https://doi.org/10.1016/j.nicl.2013.05.004.