Single-Subject Prediction: A Statistical Paradigm for Precision Psychiatry

Danilo Bzdok, Teresa M Karrer

Springer Book
Brain Network Dysfunction in Neuropsychiatric Illness: Methods, Applications and Implications

Chapter

Single-Subject Prediction: A Statistical Paradigm for Precision Psychiatry

Danilo Bzdok [1, 2, 3] & Teresa M. Karrer [1]

[1] Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, 52072 Aachen, Germany
[2] JARA, Translational Brain Medicine, Aachen, Germany
[3] Parietal Team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France

Keywords: precision medicine, prediction-inference distinction, health analytics, neuroimaging, genetics

"predictions, but not inferences, forecast what will happen"
White (1971)

1. Introduction

Psychiatric patients are treated every day by medical doctors based on clinical guidelines grounded in group definitions. Therapeutic intervention for a particular patient, however, frequently follows a trial-and-error path (cf. Rush et al., 2006). On average, only about 50% of patients benefit from a specific psychotropic drug therapy (Wong et al., 2010). Similar failure rates apply to common psychotherapeutic treatments (Hofmann et al., 2012). To render clinical care more effective, brain-imaging and genomics are among the most promising, yet expensive avenues. In research on the neural and genetic basis of psychiatric disease, the prevailing research ideology aims to discover new pathophysiological mechanisms as a stepping-stone to then reduce suffering of psychiatric patients (Insel and Cuthbert, 2015). Instead of trying to exploit newly discovered disease mechanisms towards novel treatments that help patient groups on average, an alternative research agenda is coming into reach over recent years. A fast and cost-effective strategy is to accurately predict which of the currently existing treatment options is likely to work best for one particular patient (Bzdok and Meyer-Lindenberg, 2018; Perna and Nemeroff, 2017; Stephan et al., 2017b). Increasing individualization of therapeutic intervention in precision psychiatry would also open the opportunity to automatically derive the diagnosis or expected disease course in single patients from data. Because psychiatric disorders result from disturbed brain biology, quantitative measures from *in-vivo* brain-imaging is ready for such predictive modeling approaches because structural, functional, or diffusion magnetic resonance imaging (sMRI, fMRI, dMRI) and sometimes positron emission tomography (PET) are already available in most modern psychiatric hospitals. Moreover, such spatially or temporally highly-resolved imaging techniques produce data of sufficient quantity and information granularity on which to apply data-driven statistical techniques (Eyre et al., 2016).

2. Examples of brain network phenotyping to improve predictions in single individuals

Several neuroscience studies have already demonstrated the clinical potential of combining extensive brain-imaging data with predictive data analytics towards important goals of precision medicine in psychiatry. As a first example, Drysdale and investigators (2017) were able to forecast the treatment response in a large fMRI study on patients suffering from depression (n=1188). The patients were automatically subdivided into four groups by applying data-driven clustering algorithms (hierarchical ward clustering) to their limbic and frontostriatal resting-state network data. Each of the four imaging-derived depression types was characterized by a particular pattern of neural dysfunction and could be linked to a distinct symptom constellation. The neurobiologically defined depression categories were showed to be clinically relevant as the subgroups differed in symptom reduction (according to Hamilton rating scale for depression) after a five-week treatment with

repetitive transcranial magnetic stimulation of the dorsomedial prefrontal cortex (Fig. 1). This neurostimulation method modulates connectivity of brain networks and is usually applied in refractory depression. Pattern-prediction algorithms (support vector machines) were built to predict individual treatment response based on brain connectivity data. The efficacy of treatment was correctly forecasted in nine out of ten new patients that were not involved in the previous model building procedure. Hence, a proof of concept is provided by a seminal stimulation-MRI investigation in major depressive disorder that brain-derived types of mental disease may enable the selection of the most promising treatment option in each particular patient.

This brute-force and ad-hoc character of many successful pattern-prediction approaches was shown in a neuroimaging study in aphasia to, in certain cases, also allow for insights into disease pathophysiology. Brodersen and colleagues (2011) acquired experimental fMRI data of aphasic patients and healthy controls who listened to speech and time-reversed speech stimuli (n=37). To estimate task-induced changes in directional neural coupling mechanisms, the investigators built a dynamic causal model of relevant auditory regions and the connections between these brain areas. The task-evoked thalamotemporal connectivity modulations were fed into a machine learning algorithm (support vector classification) that automatically derived rules how to discriminate disease state. The classification algorithm was able to tell apart new patients and controls almost perfectly (98% prediction accuracy). Additionally, the predictive modelling approach was directly mechanistically interpretable by revealing disease-specific brain connections in patients with aphasia (Fig. 2). This second example demonstrates that often heuristic machine-learning methods can be grounded in mechanistically meaningful neurobiological quantities to offer means for a more complete understanding of pathophysiology and predictability of brain disorders.

Our last example revisited a long-standing speculation about accelerated brain aging in schizophrenia by quantifying a scientifically valuable and clinically exploitable brain phenotype. Kraepelin himself has already suspected a neurodegenerative character of schizophrenia which is why he called the disease “dementia praecox” (Kraepelin, 1899). More recently, Koutsouleris and investigators (2014) employed predictive learning algorithms (support vector regression) to estimate brain age from the structural MRI scans of healthy subjects (n=800). The obtained trained algorithm was then used to predict brain age from brain structure of at-risk individuals who exposed sub-threshold psychotic symptoms and patients diagnosed with schizophrenia (n=230). The discrepancy between model-derived brain age and the actual age of the subjects increased from at-risk over recent onset to recurrent disease states. Moreover, the brain phenotype of a prematurely aged brain in schizophrenia was predictive of disease status in specific individuals (Fig. 3). The predictive modeling strategy could hence quantitatively formalize accelerated aging effects for successful prediction of disease state on a single-subject level, which substantiated a classical pathophysiological hypothesis in schizophrenia research.

The collection of examples exposes how the alliance of predictive modeling analytics and brain scanning can enable the individual prediction of disease state, treatment response, and clinical course in brain disorders. Against widespread belief, the fusion of “big data” and emerging statistical technologies also has the potential to open alternative windowsn on pathophysiological mechanisms in major psychiatric disorders.

3. Proposing a typology of “prediction”

Empirical research in medicine and psychology frequently invokes three different notions of how findings may allow for statements about the future of a particular individual (Bzdok, 2017a; Casella and Berger, 2002; Gabrieli et al., 2015; Woo et al., 2017):

i) Correlation, such as Pearson or Spearman’s rank correlation, computes a simple similarity metric between two series of measurements. For instance, the correlation between amygdala activity changes measured in an fMRI experiment and overall disease severity could be computed in a group of patients with schizophrenia. A high correlation between the two variables could motivate a deeper examination of the relationship between amygdala fMRI activity and disease severity. In this data-exploration setting, the goal is not to extract a model that embodies the discovered statistical relationship. As no model parameters are being

estimated in common correlation analysis, the pattern interrogated in the data cannot be directly shipped to other researchers or clinicians for application to a new individual.

ii) Linear regression approaches, such as ANOVA, ANCOVA, and MANOVA, estimates model parameters that encapsulate the mapping from explanatory (independent) input variables to explained (dependent) output variables. For instance, the linear relationship of how fMRI activity changes in the amygdala relate to the presence of schizophrenia symptoms can be extracted in a patient group. The obtained model embodies how the symptom severity can be expected to increase or decrease with a higher or lower fMRI activity in the amygdala in the subject sample at hand. The relevance of a particular explanatory variable can be quantified by comparison to the estimated importance of other explanatory variables, such as the measured fMRI activity in the prefrontal cortex, auditory cortex, fusiform gyrus, or other candidate brain regions. In this setting, a model has been generated that could be shipped to other researchers and psychiatrists for reuse in new individuals. However, for many linear-regression analyses there is a small tradition to explicitly ascertain that the isolated pattern still holds in subjects who the model has not yet seen.

iii) Single-subject prediction, such as routinely performed with support vector machines, random forests, or artificial neural-network algorithms, is typically achieved by identifying relationships in one set of subjects as a function of how these patterns persists in other individuals from a different set of subjects. Here, model parameters are typically estimated on some data while the emerging model is explicitly put to the test in some independent data from unseen individuals (Shalev-Shwartz and Ben-David, 2014). Such primary emphasis on model performance on single individuals can sometimes have the side effect of hindering understanding how the output variable is expected to change as specific input variables increase or decrease (Goodfellow et al., 2016; Hastie et al., 2001). For instance, pattern-recognition algorithms could possibly be built to derive optimal drug dosage suggestions for patients with schizophrenia from whole-brain fMRI activity, as measured from potentially tens of thousands of local brain measurements being the input variables for a given patient. In this setting, the investigator typically pursues the main goal to achieve the best-possible model accuracy for the optimal drug dosage in any single patient, even if this objective may in certain cases come at the expense of some reduced model interpretability.

In our opinion, predictions of type (iii) may be particularly tuned to the ambitions of precision medicine in psychiatry. This analysis paradigm, routinely practiced in many applications of pattern-recognition algorithms, is centered around evaluating the capacity of already extracted models to derive quantities of interest from new, potentially later encountered individuals. If an already extracted model embodying an identified relationship, reflected in the estimated parameters, is assessed in new individuals whose data were not used to estimate the parameters, the statistical analysis can be said to be an *out-of-sample prediction*. This form of building models from data has been explicitly optimized for and is naturally applicable to a single data point, such as one whole-brain scan or one sequenced genome of a particular individual. Whether an obtained model is useful in practice is judged based on its performance in achieving accurate predictions in independent individuals[1]. One may view these evaluation practices as more conservative measures when the goal is reliable single-subject predictions in patients admitted to a psychiatry hospital in the future (Bzdok and Meyer-Lindenberg, 2018; James et al., 2013). In contrast, most correlation- and linear-regression-type analyses (i and ii) reflect *in-sample predictions* where model estimation and evaluation are typically performed in the same subject sample. In this analysis paradigm, the ensuing conclusions could be closely adapted to the particular subject sample at hand. Please also note that out-of-sample prediction is distinct from a *replication* study where the same statistical approach is entirely repeated on two different subject samples, rather than extracting a model from a first set of subjects while evaluating the performance of that model in an independent second set of subjects.

[1] Niels Bohr put this point in the following words: “Prediction is very difficult, especially if it's about the future”.

4. Single-subject prediction is important for personalized medicine

We invite the reader to imagine a future of centralized medical care, where comprehensive health-related information is available on a majority of the citizens, ignoring important technical and ethical challenges for a moment (Leonelli, 2016; O'Neil, 2016). *Scientific discovery* in such rich data resources has been a primary focus of the statistical methodology traditionally used in empirical research in medicine and psychology (Bzdok, 2017b; Efron and Hastie, 2016; James et al., 2013). This modeling goal is for instance especially suited to ask, "Which gene locations *contribute to* or *are associated* with schizophrenia?" Perhaps counterintuitively, thus identified genetic risk variants may not in all cases serve best to detect *whether* a given individual is affected by schizophrenia or is healthy (Shmueli, 2010). This is because modeling for prediction typically asks a different kind of research question: "Which gene locations are *useful* to *distinguish* schizophrenic versus healthy individuals?" Finding such answers follows the heuristic agenda of prioritizing successful pattern-recognition to identify any relationship in the data that is able to derive the specified outcome in independent data, but may put smaller emphasis on *scientific insight into the neurobiological underpinnings of the schizophrenia disease*. As further example, scientific discovery inquires which genes provide mechanistic insight into explaining *why* a drug treatment for schizophrenia works in patients on average (James et al., 2013). Yet, pattern-recognition algorithms prioritize predictive genes that can reliably disambiguate *whether* a drug treatment will work in one specific patient with schizophrenia.

The data-driven identification of predictive principles from complex information has been a dominant focus in the machine-learning community (Breiman, 2001; Goodfellow et al., 2016). Pattern-recognition algorithms have frequently been employed with the goal to achieve the best possible outcome detection (Hastie et al., 2001). Typically based on modest a-priori knowledge (Abu-Mostafa et al., 2012), many algorithms in machine learning are estimated to deduce an output variable (e.g., presence of schizophrenia) from potentially many input variables (e.g., lifestyle and demographic indicators, brain function, neuropsychological tests). For qualitative output variables (e.g. healthy group versus schizophrenic group or drug responder versus non-responder), the modeling process can perform *classification*. For quantitative response variables (e.g. schizophrenia severity or probability of responding to a candidate drug), pattern-recognition algorithms are built for *regression* from potentially highly resolved input data from possibly different sources (Hastie et al., 2001). In the following, several aspects are highlighted that characterize methods embraced by machine learning with the prominent differences to tools often used in classical statistics. In particular, the distinct properties of both modeling approaches shed light on why the practices of pattern-learning methods are well suited to achieving accurate predictions *at the single-subject level*, whereas many traditional statistical approaches are frequently used in medicine and psychology to discern the "trueness" of an effect *at the group level*.

The *prediction paradigm*, such as practiced in machine learning community, departs in important ways from the *inference paradigm* grounded in classical null-hypothesis testing in types of conclusions that are to be drawn from data (Breiman, 2001; Bzdok, 2017a; Efron and Hastie, 2016). The statistical paradigms anchored at inference or prediction are common in trying to evaluate whether an effect found in some data extrapolates to another sample of observations drawn from the same underlying population (Casella and Berger, 2002; Efron, 2012). To draw statistical inference, mainstream statistics has put a major emphasis on a framework revolving around rejection of a null-hypothesis in favor of an alternative hypothesis that is contradicting the status quo. Instead, many machine-learning applications have mainly focused on building pattern-recognition algorithms that predict the discovered patterns in independent, new data that did not influence model estimation. In classical statistics, inferential conclusions are drawn by formally testing for the existence of an effect expressed under the null-hypothesis (e.g., a gene is not associated with schizophrenia) in opposition to the alternative hypothesis (e.g., a gene is associated with schizophrenia). The ensuing *p*-value indicates whether data from the subject sample at hand are too extreme to occur under the null hypothesis.

While null-hypothesis testing typically takes the form of a single-step approach, the success of precision psychiatry will probably depend on predictive models that can be extracted and then "shipped" in a two-step approach. In classical null-hypothesis testing, the p-value is computed on the *entire* data from a particular subject sample in a single process. P-values are commonly obtained from all examined individuals (in-sample) and this quantitative outcome can usually not be used to test for the *same* statistical relationship in a later encountered single individual. In contrast, methods common in machine learning can quantify the prediction performance of a previously built algorithm applied to untapped data, such as from a new incoming patient, as a performance metric and immediate practical usefulness. This process of evaluating the prediction performance of learning algorithms is typically performed by a two-step procedure called *cross-validation* (Shalev-Shwartz and Ben-David, 2014). In a first step, the machine-learning algorithm is built on a larger part of the dataset. In a second step, emerging candidate algorithms are evaluated and selected on unused data (Hastie et al., 2001). Because all conditions for independent, identically distributed observations are usually met for the left-out data, the out-of-sample prediction performance on the testing data samples can quantify how likely the same pattern could be detected in future, not yet seen patients.

It is this two-step nature of ensuring model generalization by means of cross-validation procedures that attempts to certify predictive models to be "shippable" to other mental health institutions for detection of a previously discovered pattern for a single individual (Arbabshirani et al., 2017; Stephan et al., 2017a). The option for single-subject prediction is especially relevant in precision psychiatry: the brain scans or genetic profile of a new patient can be fed into the (previously built) pattern-recognition algorithm to estimate a clinical outcome variable (Woo et al., 2017). Concretely, an association between a gene and a psychiatric disorder like schizophrenia with a statistically significant p-value does not necessarily imply that the same gene will be the best choice to successfully predict whether a given individual is affected by schizophrenia. Conversely, an effect that has been empirically shown to be highly predictive of schizophrenia disease based on cross-validation in independent individuals does not always go hand-in-hand with classical statistical tests evaluated to a significant p-value (Bzdok, 2017a; Shmueli, 2010). For these reasons, *cross-validated machine-learning algorithms and more traditional tools for null-hypothesis testing can sometimes lead to diverging conclusions in certain practical analysis settings (see Fig. 4 for an example).*

Moreover, many tools frequently used in machine learning may also be especially suited to achieve the goals of precision psychiatry because they are naturally capable of handling hundreds or thousands of outcomes at once (Bzdok and Meyer-Lindenberg, 2018; Caruana, 1998; Rahim et al., 2017). Classical null-hypothesis testing in medicine usually compares two possible output states that are expressed in the null and alternative hypothesis (Wasserstein and Lazar, 2016) - the non-preferred null-hypothesis and the alternative hypothesis posited by the investigator - with limited scaling to more possible options (Efron, 2012). The challenging question in precision psychiatry is usually not if a patient suffers from a psychiatric disorder or not. Instead, the psychiatrist rather decides which specific psychiatric disorder the patient is suffering from. Analogously, it may be clinically more relevant which treatment option a particular patient should be assigned to rather than asking whether a given patient needs a therapy or not. In fact, even when comparing a number of disease groups based on ANOVA, the null-hypothesis commonly being tested is whether all groups are equal (null-hypothesis) or not (alternative hypothesis) (Casella and Berger, 2002). For instance, one disease group being different from 9 other ones or 10 disease groups being different in every possible pair equally lead to rejection of the null-hypothesis of no difference. For these predictions of several diagnoses or treatment options in parallel, machine learning is especially suitable. This is because many machine-learning algorithms can be easily extended to the prediction of a large array of different outcomes in the same algorithm building process (Breiman and Friedman, 1997; Caruana, 1998; Rahim et al., 2017). In this way, several machine-learning approaches readily offer the opportunity to predict many clinical endpoints in a single patient. For instance, a machine-learning algorithm could be built to derive i) which disease the patient is suffering from (e.g. schizophrenia versus bipolar disorder versus major depression), ii) which disease course can be

expected (e.g. single episode versus recurrent versus chronic disease course), and iii) which treatment options will be most effective (e.g. antipsychotic drugs versus antidepressant drugs versus cognitive behavioral therapy versus their combinations).

5. Challenges to overcome

Forecasting diagnosis, disease course, and effectiveness of treatment options in incoming patients offers a promising opportunity for improving medical care for psychiatric patients. However, several challenges still need to be overcome: From a technical point of view, prediction focused statistical approaches typically require large amounts of data (Henke et al., 2016; Jordan et al., 2013). So called "big data" demand special data management skills and infrastructure that are today seldom available at clinical and research institutions (Bzdok and Meyer-Lindenberg, 2018). To further mature training of pattern-recognition algorithm, predictable patterns have been advocated to fulfill several criteria before translation into clinical practice (Woo and Wager, 2015). These authors suggested that an ideal predictive model should be (i) highly successful in diagnostic performance including both sensitivity and specificity, (ii) useful in terms of neuroscientific research, (iii) composed of precisely defined algorithms, and (iv) generalizable across different clinical settings (e.g., data acquisition means, geographic locations, and patient populations). Additionally, the application of predictive models must be checked for the economic efficiency, especially in light of the considerable costs of brain-imaging modalities such as sMRI, fMRI, dMRI and PET (Gabrieli et al., 2015). Furthermore, ethical and societal aspects will probably become an important recurring theme in discussing the future of precision medicine (Gabrieli et al., 2015; O'Neil, 2016).

6. Classical null-hypothesis testing is not obsolete

Tools for statistical hypothesis testing and more recently emerged machine learning techniques can be used to draw different types of conclusions from data. Whereas the core interest of machine-learning applications is to *predict* future events on the basis of patterns observed in data, classical statistics applications are probably more often used to *infer* scientific insight from the effects observed in data (White, 1971). Both modeling paradigms can serve distinct statistical purposes in improving psychiatric practice based on brain-imaging and genetics (Bzdok and Yeo, 2017; Shmueli, 2010). Depending on the ultimate clinical or research question, a different set of statistical tools may suggest itself as more appropriate (James et al., 2013). It is therefore important for investigators and psychiatrists to acknowledge the partly diverging modeling goals and scopes of interpretation of these two distinct statistical cultures (Breiman, 2001; Bzdok, 2017a).

Traditional null-hypothesis testing emerged in the early 20th century. This was a time in history when data were rare and expensive to acquire (Efron and Hastie, 2016; Gigerenzer, 1993). Well-controlled research experiments were carefully designed in advance. Nowadays, such datasets with few measured variables are still the norm in much research in psychology and medicine. Many early statistical tools were especially developed for such settings aiming at understanding the relationship between a few variables. If the goal is to examine whether an effect exists or which specific input variables have most impact on an output variable, classical statistics based on null-hypothesis testing is arguably still among the best tools. In practice, the focus routinely relies on the statistical analyses of few variables that tend to yield high interpretability, rather than perusing data for complex patterns that are predictive. Ideally of course, one would hope to achieve both interpretability and predictability. Several recent investigations have successfully combined "black-box" pattern-recognition analyses and model components that can be readily introspected for scientific understanding (cf. Brodersen et al., 2011).

Today, single-subject prediction becomes always more feasible due to the recent co-occurrence in data availability, computing power, and cheaper data storage (Goodfellow et al., 2016; Manyika et al., 2011). Brain-scanning and genetic measurements in psychiatry produce massive amounts of data at high granularity that classical statistical tools have not initially been invented to tackle (Efron, 2012). In contrast, machine learning was designed to extract patterns from such observational data

that was frequently acquired outside of a carefully controlled experimental context. Additionally, many machine-learning approaches specifically motivated for achieving prediction at scale, such as in thousands of individual subjects or for hundreds of outcomes, as well as when outcome variables are hard or expensive to collect. In precision psychiatry for instance, the accurate prediction of a psychiatric disease, the disease course, or efficacy of treatment options in individual patients is the relevant research goal.

However, it is important to appreciate that the potential immediate gains of the pragmatic goal to identify patterns useful to predict clinical endpoints in complex data does not preclude the longer-term urge for *understanding* the biological nature underlying psychiatric diseases like schizophrenia. Carefully designed, meticulously conducted, and logistically expansive experiments to confirm or reject a-priori verbalized research hypotheses in animals and humans will probably remain a cornerstone to generate neuroscientific insight into mental illness. As one of many potential scenarios for a happy cohabitation of the inference and prediction paradigms, pattern-recognition algorithms with successful prediction performance will probably need to undergo carefully controlled randomized clinical trials based on traditional null-hypothesis testing before regulatory stakeholders authorize translation into clinical practice.

7. Conclusions

The current diagnosis systems in psychiatry that are pervasively used every day to clinically diagnose thousands of patients based on their symptoms and to provide medical care were established largely based on expert opinion. Diagnosis *groups* have served well in clinical practice and scientific investigation to impose structure on the evasive conglomerate of mental health disorders. The now rapidly increasing amount, detail, and quality of health-related information on each given individual is announcing a major turning point in the history of psychiatry. We may increasingly abandon predefining disease concepts to then group psychiatry patients based on observed symptoms for clinical investigation of brain dysfunction. Instead, it will be increasingly possible to first quantitatively derive disease stratifications directly from brain measurements in a data-guided fashion to then capitalize on the discovered brain-based phenotypes for patient-tailored monitoring, risk assessment, and therapeutic intervention.

Approaching this future of psychiatric diagnosis systems rooted in brain biology will, in our opinion, necessitate awareness of precise notions of *prediction* and the statistical techniques with the honest capacity for deriving rigorous statements on single individuals. The perhaps most important point we tried to make in this crash course is this: Medical research has had a long-standing focus on scientific discovery revolving around understanding disease mechanisms by statistical methods targeted at establishing *statistical inference.* This statistical goal is in many cases incompatible with the pragmatic wish to somewhat blindly exploit the quantifiable consequences of brain pathophysiology to achieve most accurate *predictions* about the future of individuals based on diverse and rich biological information. Appreciation of this *inference-prediction divergence* will probably be a necessary milestone in personalized medicine research, which will ultimately benefit the well-being of suffering psychiatric patients.

Figures
Figure 1

XXX

Caption: Brain connectivity can predict treatment response in major depressive disorder. (a) Subgroups of patients clustered by resting-state connectivity profile differed in their clinical response to repetitive transcranial stimulation (rTMS) intervention on the dorsomedial prefrontal cortex (dmPFC). Treatment response rate corresponds to percentage of patients whose symptom severity was reduced by at least 25% as measured by the Hamilton rating scale for depression (HAMD). (b) Boxplot describing the distribution of symptom severity reduction in four depression biotypes. Percent improvement in symptom severity was computed as difference in total HAMD score before and after rTMS application. **P = 0.00001–0.002 (Mann–Whitney), marks significant increase compared to biotypes 2–4; *P = 0.007 (Mann–Whitney), marks significant increase compared to biotype 4. (c) Treatment responders differed in their functional coupling of dmPFC target area with other brain regions compared to non-responders: warm colors = increased connectivity, cold colors = decreased connectivity (Wilcoxon rank–sum tests, thresholded at $P < 0.005$). (d) Neuroanatomical locations of 25 brain regions (top 10%) which expose most discriminative connectivity features allowing detection of treatment responders versus non-responders. Red arrows point at rTMS target. (e) Heat maps illustrate how responders to rTMS differed in their functional connectivity profile compared to non-responders. (c-e) Brain regions colored by corresponding functional networks. Figure reused with permission from Drysdale et al. (2017).

Figure 2

XXX

Caption: Study rationale for discovering disease-relevant functional connections in aphasia disease**.** Brain activity as measured by the blood oxygen level dependent (BOLD) signal was acquired in patients with aphasia and healthy controls during a speech-processing task. In the first step, the functional data of each subject were used to estimate the parameters of a dynamic causal model (DCM). The ensuing directional connectivity estimates were supposed to capture quantities describing functional brain mechanisms. In the second analysis step, a kernel function was built to i a similarity metric between the fitted models from two subjects. The constructed kernel instantiated a model-based space of variables. In this exemplary subject, the influence of region A on region B and the self-connection of region B were particularly strong. In a third step, a machine-learning algorithm (support vector machine) was trained to disambiguate patients with aphasia and healthy controls exclusively based on connectivity strengths. In the last step, the relevance of functional coupling of brain regions A – C for discrimination of disease state could be interpreted mechanistically. In this example, the joint influences of region A on region B and C were clinically most relevant to tell apart patients and controls that used during model building (cross-validation). Figure reused with permission from Brodersen et al. (2011).

Figure 3:

XXX

Caption: Relationship between accelerated brain aging indexed by brain structure and classification of patients with schizophrenia (SCZ) versus healthy controls (HC). Tests evaluated whether the difference between actual and brain-derived age predicts disease discriminability, based on machine learning (support vector machines) in volumetric MRI data. Upper left: ROC curve of the disease classifier (blue line) and the brain age gap prediction (red line) to evaluate their performance in telling apart patients and controls not used for predictive model building. Upper right: continuous prediction (support vector regression) of patient classification values and brain age gap estimation. Both measures were highly correlated ($R2 = .53$; $T = 26.1$; $P < .001$). Lower panel: neuroanatomical overlap (yellow) of the brain regions relevant for age prediction (red) and disease state classification (green). Pattern-learning algorithms were fed by two structural brain maps: i) gray matter-Regional Analysis of brain Volumes in Normalized Space (GM-RAVENS; left) and ii) affinely registered gray matter data (right). Shared variation of patient- and age-predictive brain patterns might be an explanation for the observed correlation between brain aging and disease classification. Figure reused with permission from Koutsouleris et al. (2014).

Figure 4

XXX

Caption: Classical null-hypothesis testing and machine-learning algorithms can lead to diverging conclusions. The toy data distributions may reflect two groups (e.g. patients with schizophrenia versus healthy controls) as assessed by differences in a particular brain measurement (e.g. amygdala activity). Group differences are evaluated by i) null-hypothesis testing using two-sample t-tests ("P-value") and ii) machine-learning using a classification algorithm to predict which group each brain data point belongs to as indicated by dotted rule (“Classification”). In three cases with different brain data distributions, (A) t-test was statistically significant, while classification accuracy was poor. (B) Based on this data scenario, t-test was not statistically significant, while classification accuracy was high. (C) In yet another point distribution, both t-test was statistically significant and classification accuracy was high. This artificial example illustrates that null-hypothesis testing and machine-learning algorithms constitute two different statistical cultures that do not necessarily judge data distributions by the same aspects of evidence. Hence, group effects as assessed by significant p-values do not always entail a high prediction performance, and vice versa. Figure reused with permission from Arbabshirani et al. (2017).

References

Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H.T., 2012. Learning from data. AMLBook, California.

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.

Breiman, L., 2001. Statistical Modeling: The Two Cultures. Statistical Science 16, 199-231.

Breiman, L., Friedman, J.H., 1997. Predicting multivariate responses in multiple linear regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59, 3-54.

Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput Biol 7, e1002079.

Bzdok, D., 2017a. Classical Statistics and Statistical Learning in Imaging Neuroscience. Front Neurosci.

Bzdok, D., 2017b. Classical Statistics and Statistical Learning in Imaging Neuroscience. Frontiers in Neuroscience 11, 543.

Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: Opportunities and challenges. Biological Psychiatry: CNNI, in press.

Bzdok, D., Yeo, B.T.T., 2017. Inference in the age of big data: Future perspectives on neuroscience. Neuroimage 14, 549-564.

Caruana, R., 1998. Multitask learning. Learning to learn. Springer, pp. 95-133.

Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Casey, B.J., Dubin, M.J., Liston, C., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nature Medicine 23, 28-38.

Efron, B., 2012. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press.

Efron, B., Hastie, T., 2016. Computer-Age Statistical Inference. Cambridge University Press.

Eyre, H.A., Singh, A.B., Reynolds, C., 2016. Tech giants enter mental health. World Psychiatry 15, 21-22.

Gabrieli, J.D., Ghosh, S.S., Whitfield-Gabrieli, S., 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85, 11-26.

Gigerenzer, G., 1993. The superego, the ego, and the id in statistical reasoning. A handbook for data analysis in the behavioral sciences: Methodological issues, 311-339.

Goodfellow, I.J., Bengio, Y., Courville, A., 2016. Deep learning. MIT Press, USA.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer Series in Statistics, Heidelberg, Germany.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., Sethupathy, G., 2016. The age of analytics: Competing in a data-driven world. Technical report, McKinsey Global Institute.

Hofmann, S.G., Asnaani, A., Vonk, I.J., Sawyer, A.T., Fang, A., 2012. The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. Cognitive Therapy and Research 36, 427-440.

Insel, T.R., Cuthbert, B.N., 2015. Brain disorders? Precisely. Science 348, 499-500.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. Springer.
Jordan, M.I., Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, National Research Council, 2013. Frontiers in Massive Data Analysis. The National Academies Press, Washington, D.C.
Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rossler, A., Moller, H.J., Reiser, M., Pantelis, C., Meisenzahl, E., 2014. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophrenia Bulletin 40, 1140-1153.
Kraepelin, E., 1899. Psychiatrie. Ein Lehrbuch für Studierende und Ärzte, 6 ed. Barth, Leipzig.
Leonelli, S., 2016. Data-centric biology: a philosophical study. University of Chicago Press.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A., 2011. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute.
O'Neil, C., 2016. Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy, New York: Crown.
Perna, G., Nemeroff, C.B., 2017. Personalized Medicine in Psychiatry: Back to the Future. Personalized Medicine in Psychiatry 1, 1.
Rahim, M., Thirion, B., Bzdok, D., Buvat, I., Varoquaux, G., 2017. Joint prediction of multiple scores captures better individual traits from brain images. Neuroimage 158, 145-154.
Rush, A.J., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., Niederehe, G., Thase, M.E., Lavori, P.W., Lebowitz, B.D., 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. American Journal of Psychiatry 163, 1905-1917.
Shalev-Shwartz, S., Ben-David, S., 2014. Understanding machine learning: From theory to algorithms. Cambridge University Press.
Shmueli, G., 2010. To explain or to predict? Statistical science, 289-310.
Stephan, K.E., Schlagenhauf, F., Huys, Q.J.M., Raman, S., Aponte, E.A., Brodersen, K.H., Rigoux, L., Moran, R.J., Daunizeau, J., Dolan, R.J., 2017a. Computational neuroimaging strategies for single patient predictions. Neuroimage 145, 180-199.
Stephan, K.E., Schlagenhauf, F., Huys, Q.J.M., Raman, S., Aponte, E.A., Brodersen, K.H., Rigoux, L., Moran, R.J., Daunizeau, J., Dolan, R.J., 2017b. Computational neuroimaging strategies for single patient predictions. Neuroimage.
Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. Am Stat 70, 129-133.
White, A.R., 1971. Inference. The Philosophical Quarterly 21, 289-302.
Wong, E.H.F., Yocca, F., Smith, M.A., Lee, C.-M., 2010. Challenges and opportunities for drug discovery in psychiatric disorders: the drug hunters' perspective. International Journal of Neuropsychopharmacology 13, 1269-1284.
Woo, C.-W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017. Building better biomarkers: brain models in translational neuroimaging. Nature Neuroscience 20, 365-377.
Woo, C.W., Wager, T.D., 2015. Neuroimaging-based biomarker discovery and validation. Pain 156, 1379-1381.