Emerging Shifts in Neuroimaging Data Analysis in the Era of "Big Data"

6

Danilo Bzdok, Marc-Andre Schulz, and Martin Lindquist

Advances in positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) have revolutionized our understanding of human cognition and its neurobiological basis. However, a modern imaging setup often costs several million dollars and requires highly trained technicians to conduct data acquisition. Brain-imaging studies are typically laborious in logistics and data management, and require costly-to-maintain infrastructure. The often small numbers of scanned participants per study have precluded the deployment of and potential benefits from advanced statistical methods in neuroimaging that tend to require more data (Bzdok and Yeo 2017; Efron and Hastie 2016). In this chapter we discuss how the increased information granularity of burgeoning neuroimaging data repositories—in both number of participants and measured variables per participant—will motivate and require new statistical approaches in everyday data analysis. We put particular emphasis on the implications for the future of precision psychiatry, where brain-imaging has the potential to improve diagnosis, risk detection, and treatment choice by clinical-endpoint prediction in single patients. We argue that the statistical properties of approaches tailored for the data-rich setting promise improved clinical translation of empirically justified single-patient prediction in a fast, cost-effective, and pragmatic manner.

D. Bzdok (✉)
Department of Psychiatry and Psychotherapy, RWTH Aachen University, Aachen, Germany

Jülich Aachen Research Alliance (JARA)—Translational Brain Medicine, Aachen, Germany

Parietal Team, INRIA, Gif-sur-Yvette, France
e-mail: danilo.bzdok@rwth-aachen.de

M.-A. Schulz
Department of Psychiatry and Psychotherapy, RWTH Aachen University, Aachen, Germany

M. Lindquist
Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA

I. C. Passos et al., *Personalized Psychiatry*,
https://doi.org/10.1007/978-3-030-03553-2_6

6.1 Blessing and Curse of Increasing Information Content in Neuroimaging

The notion of "big data" in modern neuroimaging arises in two related, yet importantly different ways. On the one hand, the number of observed variables per participant, called "feature dimensionality" (p) and, on the other hand, the available "sample size" (n) of scanned participants. In traditional experimental studies in psychology, neuroscience, and medicine the number of observed variables has rarely exceeded the number of participants. Concretely, many common neuropsychological questionnaires and medical assessments capture <30 items—few in comparison to the often hundreds of participants in clinical trials. This so-called "long-data" setting (participants $n >$ variables p) is the realm of classical statistics. Around the turn of the century, the development of whole-genome sequencing and brain-imaging led to biology and medicine entering the high-dimensional, or "wide-data", setting (variables $p >>$ participants n; Efron 2012; Efron and Hastie 2016). For example, in genetics, the feature dimensionality from the ~3 billion base pairs or the >100,000 single nucleotide polymorphisms summarizing the human genome vastly exceeds the size of typically collected participant cohorts.

The brain sciences have recently been argued to be the most data-rich among all medical specialties (Nature Editorial 2016). A single brain scan with high-resolution MRI can easily exceed 100,000 variables that collectively describe brain morphology or a type of neural activity. However, over the last 20 years, the sample size in a typical brain-imaging study has rarely exceeded 50–100 participants. We argue that important statistical consequences arise from the divergence of the "n-p ratio" (the relation between the number of participants and the number of variables per observation) in the classical and high-dimensional settings.

High-resolution MRI increases the potential for new neurobiological findings, but the increased information detail in the brain recordings also exacerbates the dangers of the so-called "curse of dimensionality" (Bellman 1957; Friedman et al. 2001). Humans are accustomed to operating in the physical world and our geometric perception is fine-tuned to 3-dimensional environments. Human intuition regarding geometric properties, such as volume or distance, tends to struggle and eventually go awry in high-dimensional spaces. Mathematically, an increase in feature dimensionality (imagine going from a line to a square to a cube) leads to an exponential increase in the input-data space, and the available data points become increasingly sparse so that even the volumetric brain scans of monozygotic twins may look dissimilar in high dimensions. In brain-imaging, an increase in resolution (such as more voxels or more scans per time) will offer more detailed information, but the higher information granularity will also make the relevant neurobiological structure more difficult to identify. With respect to the brain data themselves, this volume increase entails that, with each (uncorrelated) new variable, investigators would potentially need to scan exponentially more participants to populate the input variable space at the same density (Bishop 2006). With respect to machine learning algorithms applied to brain data, it means that with more input variables per participant, a pattern-recognition algorithm will increasingly struggle to find

interesting statistical relations that exist in the data. The considerable increase in data abundance and complexity will put many classical statistical methods at risk of being deemed obsolete, and replaced by modeling approaches better tailored to the new data reality in imaging neuroscience.

6.2 Recent Trends for Data Collection and Collaboration Across Laboratories

The acquisition of brain-imaging data at scale is a challenging undertaking due to a variety of technical, logistic, and legal factors. These hurdles range from the need for time-effective and harmonized measurement protocols, to the participants' informed consent for sharing their data. New brain-imaging projects have tackled many of these challenges and aim to provide general-purpose datasets to the neuroscientific and psychiatric research community. Here, we give an overview of the current state of "big-data" brain-imaging, and illustrate important ramifications for data-analysis practices due to the increasing data accumulation.

Three data initiatives stand out in the brain-imaging landscape (Smith and Nichols 2018): The Human Connectome Project (HCP), the UK Biobank (UKBB) Imaging Study, and the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) Consortium. The HCP, launched 2009, was one of the earliest attempts to create a rich reference dataset for the brain-imaging community. As the name suggests, an important goal of the HCP initiative was to promote insight into functional connectivity architecture by providing extensive multimodal data on a large number of healthy participants. The HCP consortium recently completed multi-modal measurements of over 1200 healthy adults (aged 22–35), including 300 twin pairs. For each participant, the project gathered structural, functional, and diffusion MRI, genotyping data, as well as a large variety (>400) of demographic, behavioral, and lifestyle indicators. With genetic profiling and extensive phenotyping with several thousand descriptors, UKBB is even more comprehensive. This data collection initiative set out in 2006 to gather genetic and environmental (e.g., nutrition, lifestyle, medications) data from 500,000 volunteers, and is currently the world's largest biomedical dataset. UKBB recruited adults between the ages of 40 and 69. The participants will be followed for >25 years, including repeated measurements and access to their electronic health records. In 2014 UKBB launched its brain-imaging extension, aiming to gather structural, functional, diffusion, and susceptibility-weighted MRI of 100,000 participants by 2022 (Miller et al. 2016). Yet another ambitious attempt to create a large-scale neuroimaging dataset is the ENIGMA consortium, launched in 2009. Compared to UKBB and HCP, ENIGMA takes a different approach by centrally coordinating research projects and providing each participating group with analysis pipelines and quality control protocols. The software is run independently by each acquisition site and the ensuing results are combined into integrative summary analyses, possibly across different imaging modalities (i.e. structural, functional, or diffusion MRI). Because of this, the sample size can be on the order of several thousand participants depending on the availability of brain-scans directly relevant for a particular research question.

In sum, we portrayed three contemporary data-aggregation projects, which have substantially different research agendas. While UKBB is above all a medical dataset and was designed for large-scale population epidemiology, the ambition of HCP lies in functional and anatomical connectivity in healthy subjects, whereas ENIGMA has an important emphasis on genetic profiling in combination with brain scanning. Many more comparable datasets are in the making and should, within the next decade, multiply the amount of brain imaging data available for research.

Compared to many traditional MRI experiments consisting of only a few dozen participants, large-scale projects such as HCP and UKBB have unprecedented strengths and pave the way for new neuroscientific insights. A key aspect is the study design. Most imaging studies have a *retrospective* or *cross-sectional* nature in that the investigators first decide what they are looking for (e.g., a certain disease diagnosis or behavioral facet), and then recruit participants that fulfill the inclusion criteria. The phenotype of interest has already been identified, and the study is in some sense looking into the past. In contrast, UKBB is a *prospective* epidemiological study. A broad sample of the population is included in the expectation that a relevant set of the participants will experience a variety of health-relevant events at some point in the future. For example, among the 100,000 participants to be brain-scanned, ~1800 are expected to develop Alzheimer's disease by 2022, ~8000 will develop diabetes, ~1800 will have experienced a stroke, and ~1200 will be affected by Parkinson's disease (Sudlow et al. 2015). Once these medical conditions have developed, data will be available to the investigators consisting of information before, and on the path to, disease onset. This potentially unprecedented wealth of longitudinal information can be leveraged to identify early disease markers and new risk factors; perhaps even chart hypotheses that might not have occurred to researchers when designing a retrospective study. As most diseases only develop in a small percentage of the population, sampling a large number of participants is necessary for prospective studies to gain traction. Such future-oriented data aggregation designs have great potential for early disease detection and transdiagnostic stratification in mental health.

Despite much enthusiasm, the creation, curation, and collaboration of extensive brain-imaging datasets also raise a series of technical challenges (Arbabshirani et al. 2017; Bzdok and Meyer-Lindenberg 2018; Woo et al. 2017). Inter-scanner differences and the need for quality control at scale come into play. Effective data collection is complicated by the fact that brain-imaging is highly sensitive to differences in scanner type and configuration. For example, scanner-specific differences in the measured longitudinal changes in regional gray matter volume emerge even for identical scanner models (Takao et al. 2013). Multi-site data collection projects should take into consideration that these inter-scanner differences can confound statistical analysis (Focke et al. 2011). Reducing the heterogeneity of the acquired data is either costly (i.e., requires multiple identical setups), or reduces collection efficiency (i.e., single-scanner bottleneck). Different existing projects make different trade-offs between collection efficiency and incurred inter-scanner effects. ENIGMA prioritizes collection efficiency by working in parallel on a variety of different types of scanners. To minimize confounding influences due to inter-

scanner effects, UKBB uses identical scanner hardware at the different acquisition sites, while the HCP has relied on a single scanner for the entirety of their data acquisition.

Moreover, common quality control procedures that are usually performed by hand can become infeasible. Undetected technical artifacts, movement artifacts, or human error in applying the measurement protocol can distort statistical analysis. In traditional small- to medium-scale studies, even in HCP, it was still possible to perform quality control manually. A researcher or technician could visually inspect the data for each participant and scanning modality to check for errors and artifacts. The sheer amount of brain data that is generated in large-scale brain-imaging projects makes the manual approach to quality control overly time-consuming. UKBB has conceived and implemented automated quality control procedures (Alfaro-Almagro et al. 2018). This approach uses pattern-learning algorithms to model the data distribution and automatically identify artifacts and measurement errors. UKBB, HCP, and ENIGMA have invested in elaborate automated processing pipelines and protocols to detect and correct errors and guarantee standardized data.

6.3 Anticipating Upcoming Shifts in Statistical Practice

Once successfully collected and controlled for quality, massive brain-imaging datasets allow for more ambitious statistical analyses than standard studies consisting of only a few dozen participants. Recently, more advanced statistical and computational approaches have emerged to address new research goals, such as the search for neuroimaging biomarkers and hidden brain phenotypes that are demonstrated to be useful at the single-subject level. We will discuss in detail four key directions in which the increased amount of data in brain-imaging is likely to usher in changes to everyday statistical data-analysis practice. We anticipate, first, a trend for parametric methods to be complemented by flexible non-parametric methods that allow for more detailed models of the brain. Second, a trend for discriminative methods to be complemented by more applications of generative models that aim to uncover the mechanisms for how the observed data arose. Third, a tendency for frequentist and Bayesian approaches to be combined for data analysis solutions that are both computationally cheap and holistic in interpretation. Fourth, out-of-sample generalization will become an increasingly attractive alternative to classical null-hypothesis hypothesis testing. Below, we discuss each direction in turn. We will also describe how "big-data" innovations can potentially aid in the analysis at the single-subject level, providing a mechanism for precision psychiatry.

An important benefit of large-scale data collection is that it allows for more *expressive* models for describing phenomena in the brain—models that can capture higher-order non-linear interactions in the data and are able to represent more subtle aspects about the brain (i.e., increased model expressiveness). There are two ways in which this can happen. First, increased participant sample sizes make it possible to extract details and nuances from the data distribution that would be indistinguishable from random fluctuations in small studies. Second, more data points allow for a

higher number of parameters to be reliably estimated, allowing for more expressive models that can instantiate more complicated neural phenomena (i.e., models that can reproduce potentially extremely complex statistical relationships; Devroye et al. 1996; Bickel and Doksum 2007).

Classical statistical methods, such as *t*-test, analysis of variance (ANOVA), and linear regression, used for example in the widely distributed statistical parametric mapping (SPM) software package, do not exhibit the properties necessary for representing increasingly complicated brain properties with an increasing number of participants. Classical methods attempt to model data with a fixed, limited number of parameters, and usually make rigid assumptions about the underlying structure of the brain measurements. For example, the *t*-test and ANOVA usually assume Gaussianity regardless of the underlying data distribution observed in the MRI brain scans. After accumulating enough participants to detect a statistically significant effect, additional data may yield little additional insights. In fact, classical methods may frequently underfit the data in more complex data settings with many input variables. The use of a fixed number of parameters qualifies these methods as *parametric*. In contrast, *non-parametric* approaches (Fig. 6.1) typically make weaker assumptions about the underlying structure of the acquired brain data. Here the number of parameters can flexibly adapt with the number of participants, and is potentially infinite. Data from more participants allow for more nuanced quantitative brain representations, based on less rigid statistical models.

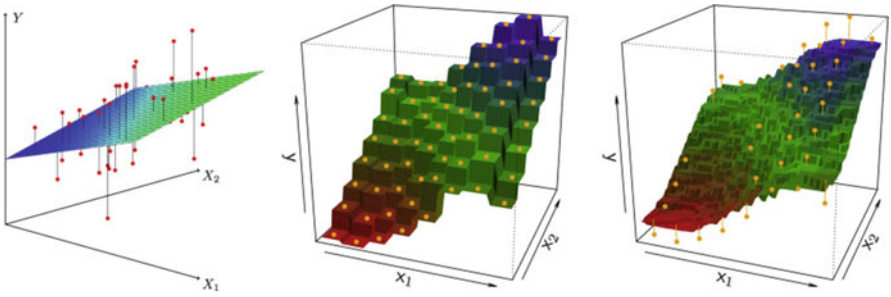

Fig. 6.1 *Parametric vs. non-parametric approaches.* Non-parametric methods (with a number of parameters that scales with increasing data availability) are more flexible than parametric methods (with a fixed number of parameters). We illustrate this distinction for the case of predicting a target variable Y based on two input variables X_1 and X_2. The parametric method of linear regression (left) always estimates three parameters defining the plane that best explains variation in the data. The number of parameters is independent of the number of data points and independent of the shape in which the data points are distributed—the end result is always a plane. In contrast, the non-parametric k-nearest-neighbor algorithm (middle and right) can adapt to a more complex shape by increasing the number of parameters in step with the number of available data points. With ample amount of available data points (right, $k = 9$), the shape of the regression surface turns from a coarse step function (middle, $k = 1$) into a smooth approximation of the data distribution (right). Non-parametric methods adapt their number of parameters in step with the number of data points and can thus reproduce more complex shapes and distributions. Reproduced from James et al. (2013)

An example of a non-parametric method is the *k*-nearest neighbor algorithm (Fig. 6.1). A sample (e.g., a T1 image of a healthy or schizophrenic participant) is classified by the class membership (disease status) of the majority of its closest data points in the dataset (the other participants). As the number of samples increase, more details of the data distribution (e.g., individual brain anatomy) can be captured leading to a more refined quantitative representation of the brain phenomenon under study. Other popular examples of non-parametric methods are decision trees (and tree-based methods such as random forests) and kernel support vector machines. In both approaches the number of model parameters scales naturally with the number of participants. Extensive biomedical datasets are ideal for using non-parametric methods to capture previously unobserved neurobiological properties that might be ignored when using parametric methods alone.

An example of the application of non-parametric methods in brain-imaging is the investigation by Gennatas et al. (2017) on how gray-matter changes with age in a large neurodevelopmental dataset (Pennsylvania Neurodevelopmental Cohort, 1189 participants aged 8 to 23). A parametric approach would have been to use an instance of the (parametric) generalized linear model (GLM) to relate MRI gray-matter measures to age, that is to estimate coefficients for the variables (gray-matter measures) that best predict the target (age). Instead, Gennatas and colleagues used a non-parametric extension of the GLM called "generalized additive models" (GAM; Hastie and Tibshirani 1990). Instead of fitting a coefficient for each input variable, GAMs estimate an adaptive functional form linking each individual variable with the respective output variable. With more data points (participants), the identified arbitrarily complex input-output functions could more accurately reflect the interaction between gray matter voxels and overall participant age. The GAM is thus able to describe and exploit highly non-linear statistical relationships to which the GLM would be blind[1]. Integrating the non-linear relationships between regional gray-matter volumes and age increased the goodness of fit of the model, leading to less noisy parameter estimates and therefore to enhanced understanding of gray-matter changes in individual brain regions across the lifespan.

As a second important distinction, statistical models can be used to address a research goal directly—discriminative models—or additionally learn intrinsic structure from the data at hand—generative models (Fig. 6.2). As an analogy, assume somebody wants to distinguish between speech from Japanese and Chinese speakers. A generative model would first try to learn the grammar, vocabulary, and phonology of both languages. Only then would the model address the classification-goal of disambiguating whether a certain speaker is Japanese or Chinese based on an explicit internal representation of what each of the two languages looks like. A discriminative model, on the other hand, would use any aspect of the speech, such as the intonation or the frequency of certain phoneme combinations, to somewhat blindly distinguish the speaker groups—even if no deeper understanding is obtained

[1]The only way for the GLM to describe non-linear interactions is to anticipate the particular effect and introduce the corresponding higher-order terms explicitly into GLM model from the beginning.

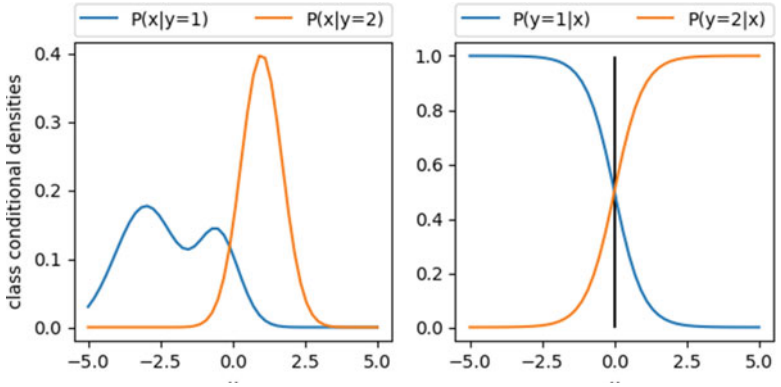

Fig. 6.2 *Generative vs. discriminative approaches.* Patients (black) and controls (red) both undergo the same biomedical evaluation. The result of the test is indicated on the x-axis, the likelihood that a participant of either class will receive a particular result is indicated on the y-axis (left). There exist two statistical approaches to predict if a given participant is patient or control based on the test result. A discriminative model (right) estimates a decision boundary (vertical line) that optimally separates the patients from the controls. Apart from the decision boundary, no other information is extracted from the data. A generative model (left) estimates the full probability distributions of both the patient and control group. The probability distributions are then used to determine whether a given participant is more likely to be patient or control. The generative model also captures information about the data distribution that does not directly help to distinguish patients from controls (e.g., information about the far ends of the probability distributions or about the density bump at $x = -1$). This "unnecessary" information can reveal important biological insights: In this case, the density bump at $x = -1$ could indicate that the patient group is in fact composed of two different groups with distinct symptom profiles. Inspired by Murphy (2012)

about the speech's content and structure. In a large number of application domains in empirical research, discriminative models have dominated statistical analysis. In the example of distinguishing[2] a healthy group from a schizophrenic patient group, discriminative models (e.g., logistic regression, support vector machines) learn a decision boundary between the participants from each group (think of a dividing line between categories, e.g., healthy vs. diseased)—or, more formally, they estimate the posterior probability[3] $P(y|x)$, without extracting an explicit representation of each class to be distinguished. In contrast, generative models (e.g., naive Bayes classifier) estimate the joint distribution $P(x,y)$—or, more informally, generative

[2]The classification setting serves as an illustration only. Discriminative methods exist independently of the classification—regression divide. For example, the clustering algorithm k-means is discriminative in the sense that it finds decision boundaries between clusters, although it attempts neither classification nor regression.

[3]$P(y|x)$ is the so-called conditional (in the Bayesian terminology the "posterior") probability: The probability of an event y (e.g., the patient is diseased) under the condition that another event x (e.g., a certain brain anatomy measured by MRI) has already occurred. $P(x,y)$ is the so called joint probability: The probability of x and y occurring together.

methods model the process by which the data was generated (Jebara 2012; Bishop and Lasserre 2007). The class posterior distributions $P(y|x)$ can then be derived using Bayes' rule.

Importantly, generative models have the intrinsic ability to produce new, artificial data samples. This ability to create never-observed data that is characteristic for one of the classes has an appealing advantage. Sampling from the generative model and visually inspecting the generated samples can provide direct insights into the inner workings of the brain phenomenon under study. In a model of the brain, where one model parameter is hypothesized to represent age, varying this parameter would allow the investigator to see a brain aging before their eyes—providing insight into age-related brain changes. However, a natural caveat is that the results will only be as good as the underlying model. If the model does not accurately depict the phenomena in question, the output of a generative model will be similarly flawed.

As a consequence, generative models are usually easier to interpret than most discriminative models because the modeled internal representation of what the data "looks like" (i.e., the conditional variation between input variables, output variables, and possible hidden variables) has been noted to capture biologically meaningful structure in previous brain-imaging studies. Furthermore, many generative models work by adaptively modeling hidden states of a system, or by finding a compact set of hidden factors that describe the dynamics of the system at hand. This process is often called *latent factor* discovery (Goodfellow et al. 2016, Chap. 13). A compact set of latent factors is usually easier to interpret than potentially high-dimensional brain-imaging input data (Fig. 6.3). A simple example of such a latent factor based generative model is the commonly used independent component analysis (ICA). ICA reduces the data to a manageable number of hidden directions of variation. As a generative model, ICA is able to produce never observed, artificial data samples based on the extracted latent factors. Such sources of variation underlying the observations can be easily interpreted (e.g., by plotting which brain areas associated with which latent factor) and can uncover previously unknown information about the brain in both health and disease. Given enough samples of resting-state fMRI time series, ICA is able to both find hidden multivariate patterns that together explain the variation in the data (e.g., the default mode network) and generate new artificial brain images from the derived factors. The combined statistical goal of generative methods to model hidden states of the brain phenomena and minimize an optimization criterion at hand (e.g., prediction performance) is usually more challenging than the statistical goal of discriminative models to simply find a decision boundary between classes. This explains why generative models tend to require brain data from more participants and why they are now becoming increasingly attractive with large-scale datasets.

A common generative model in brain-imaging is dynamic causal modeling (DCM) invented by Friston et al. (2003). The goal of DCM is to estimate directed "effective connectivity", that is, the functional influence that one brain region exerts on another brain region. DCM explicitly estimates interactions between neuronal populations in the context of a biophysical model of the hemodynamic response. This characteristic makes DCM a generative model with neurobiological plausibility

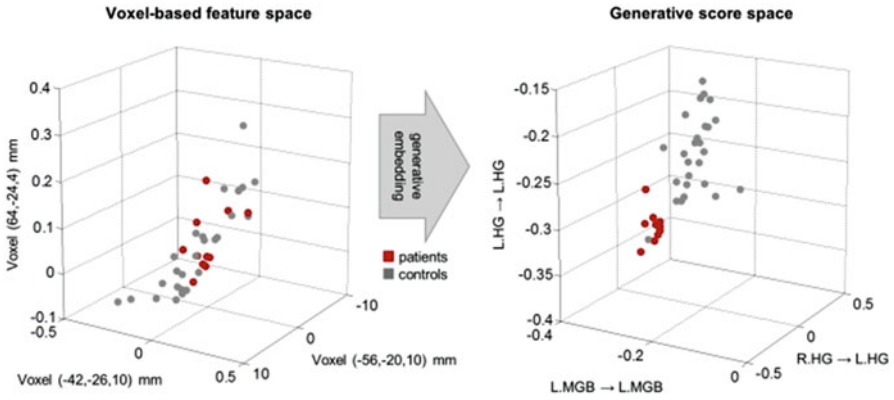

Fig. 6.3 *Latent factor model in action.* Dynamic causal modeling is a brain-imaging analysis technique that can be used to model the functional connectivity in the brain. DCM uses fMRI activity data to estimate the degree of connectedness between predefined brain regions. The DCM model parameters can be seen as a different perspective on the same data: Each participant has different fMRI activity and thus different estimated DCM model parameters. Here, whole-brain fMRI data do not lend themselves to distinguish patients from controls. The figure on the left shows how patients and controls are distributed in the space spanned by three voxels ("voxel-based feature space"). The DCM parameters capture more meaningful biological concepts than individual voxels, and patients and controls become separable. The figure on the right shows how participants form clusters of patients and controls when viewed in the space spanned by three DCM connectivity parameters ("generative score space"). Reproduced from Brodersen et al. (2011)

that is able to synthesize plausible hemodynamic activation patterns from hidden neural activity in brain regions. In addition to various human fMRI studies, the plausibility of DCM has been directly evidenced in rats by successfully relating intracerebral EEG recordings to rat fMRI (David et al. 2008).

It should be noted that not every generative model is based on latent factor discovery, and not every latent factor model qualifies as a generative model. Some generative approaches work by transforming random input vectors (e.g., generative adversarial networks) or autoregressive models (e.g., pixelRNN, waveNet) and do not lend themselves to easy introspection of the underlying statistical relationships by the investigator. An example of a non-generative latent factor model is classical canonical correlation analysis[4] (CCA). This exploratory method is similar to principal component analysis in that it reduces the data to orthogonal principal vectors, but instead of maximizing explained variance, CCA maximizes the correlation between two (lower-dimensional) latent factors of two data "views", for example, brain-imaging on the one hand and behavioral performance scores on the other hand. CCA thus identifies aspects of brain-imaging data and behavioral data that exhibit maximal linear correspondence with each other.

[4]Although there exists a generative probabilistic variant of CCA, the widely used classical CCA is not inherently generative.

For instance, Wang et al. (2018) used canonical correlation analysis to provide some of the first evidence for distinct neurobiological underpinnings of different subjective experiences of mind-wandering. Such stimulus-independent cognitive processes are associated, amongst others, with executive performance and creativity indicators. To provide evidence that mind-wandering is not a homogeneous psychological construct, but instead comprises a range of cognitive architectures and functions, the authors employed CCA with resting-state fMRI data as one view and self-reported experience, cognitive performance, and psychological well-being data as the other view. The CCA revealed latent factors that simultaneously described individual variation in self-reported experience and connectivity in the default mode network, as well as factors uniquely related to aspects of cognition, such as executive control and creativity. These findings, enabled by the unique modeling capabilities of CCA, provided evidence that distinct brain dimensions collectively contribute to different cognitive aspects underlying the mind-wandering experience.

Traditionally, perhaps the most important distinction in statistics in general and in neuroimaging in particular has been between *frequentist* and *Bayesian* models (Freedman 1995). To illustrate, let us consider the example of medical research. A Bayesian researcher would happily introduce prior knowledge from past research and experience into her statistical inferences to guide further upcoming research. These a-priori assumptions placed on the model parameters in combination with Bayes' rule yield full probability distributions, that is, a point estimate and detailed information on the uncertainty that comes with, for example, the effectiveness of the proposed treatment. The frequentist medical researcher, on the other hand, would shy away from the subjectivity of making a-priori assumptions before studying the data. She obtains an estimate without detailed uncertainty information—for the treatment effectiveness that hold with fewer assumptions about the underlying data-generating process. Intuitively, Bayesian statistics is a good choice for several research questions being asked using neuroimaging techniques. Commonly accepted knowledge of brain anatomy and physiology can for instance be used as a basis to come up with a-priori assumptions that guide the model fitting process. In the example of DCM, interactions between neuronal populations are modeled not just based on the experimental data, but instead the modeling process is couched in probabilistic a-priori knowledge concerning hemodynamic parameters, anatomical regions, and more.

In contrast to many approaches to full Bayesian inference, performing statistical data analysis using a frequentist approach is usually computationally cheaper (Bishop and Lasserre 2007; Jordan 2011; Yang et al. 2016). The "model evidence" term in Bayes' formula is typically the source of the much increased computational load in the Bayesian setting. It is an integral over all possible values of all relevant parameters (which are often much more numerous than the feature dimensionality of the actual quantitative observations in the brain) that usually cannot be directly solved, and even reaching approximate solutions is computationally challenging in many cases. A common tool for these approximations, the family of Markov chain Monte Carlo (MCMC) methods, is an iterative algorithm that is not easily parallelizable. These hurdles become even more severe in domains such as brain-

imaging, where an arms race for increasingly finer spatial and temporal resolution is constantly pushing the feature dimensionality of the brain scans. One potential solution to the computational expense of Bayesian inference in many applications to extensive brain data is the integration of Bayesian and frequentist modeling paradigms. An example of such a hybrid approach is variational inference—a widespread modeling solution to approximate complicated Bayesian integrals (Jordan et al. 1999). Another hybrid approach that has been shown effective is *shrinkage*, a statistical estimation method in which individual observations "borrow strength" from a larger group of observations (Bzdok et al. 2017; Varoquaux et al. 2010; Mejia et al. 2015). Shrinkage is implicit in Bayesian inference, penalized likelihood inference, and multi-level models and is directly related to the empirical Bayes estimators commonly used in neuroimaging (Friston et al. 2002; Friston and Penny 2003).

A combined Bayesian-frequentist approach was also applied by Brodersen et al. (2011) in the aim of computational psychiatry. Faced with the challenge of classifying a small number of participants into healthy and diseased groups based on the high-dimensional input data from all voxel activities in the whole fMRI time series, they introduced classification via "generative embeddings". These investigators used Bayesian, generative dynamic causal modeling to compute effective-connectivity models for each participant. The DCM model parameters were then used as a low-dimensional effective summary of the high-dimensional voxel data (Fig. 6.3). This dimensionality reduction via domain knowledge (i.e., priors on brain anatomy and physiology in the DCM) mitigated the curse of dimensionality and, in a subsequent step of the modeling approach, allowed for the data to be classified by a frequentist support vector machine, thereby combining the strengths of both Bayesian and frequentist inference.

Finally, in mainstream statistics as routinely applied in medicine, psychology, and brain-imaging, new knowledge is typically derived from data by means of *null-hypothesis testing*, that is testing whether or not an observation is too extreme to be plausible under the null-hypothesis of no effect (Fisher and Mackenzie 1923; Neyman and Pearson 1933). In a drug trial, the null-hypothesis would be that the new drug is no more effective than a current standard treatment. A measured effectiveness that defies explanation as a random fluctuation in the experiment would lead the investigator to discard the null-hypothesis and establish the superiority of the new drug. An overarching theme of classical statistics in the twentieth century was to optimally exploit small sample sizes using low-dimensional parametric models (Efron and Hastie 2016).

The recent advent of large-scale data collection has had two important consequences. First, caveats emerge for hypothesis testing in ever more high-dimensional neuroimaging data. The "multiple comparisons" problem becomes increasingly challenging to address in the wide-data scenario (Miller 1981; Efron 2012). The traditional approach in the brain-imaging community is called "mass univariate" analysis and performs separate statistical tests for each brain location. When many null-hypothesis tests are being carried out in concert, an increasing number of false positive findings will plague the data analysis and subsequent interpretation. Many

commonly used methods to explicitly account for the number of false positives, such as Bonferroni's method for family-wise error correction, work by increasing the threshold for statistical significance in a conservative fashion, which substantially increases the number of participants whose brain data are necessary to reject a given null-hypothesis.

On the other hand, if the number of variables is small (e.g., after reducing whole-brain data to a lower-dimension using independent component analysis) but the number of participants happens to be much larger, even very small, practically irrelevant statistical effects will sooner or later become significant (Berkson 1938). For instance, brain-behavior correlations of $r \approx 0.1$ were consistently found to be statistically significant when considering a sample of $n = 5000$ participants even after correction for multiple comparisons (Miller et al. 2016). This and similar examples illustrate that, in the era of "big-data" neuroimaging, hypothesis testing may more and more often struggle to distinguish between statistical and practical significance. In sum, the traditional null-hypothesis testing frameworks may have to tackle new difficulties in analysis settings with a lot of input variables ("wide-data" or $n << p$ setting) and when brain data from a large human population are considered ("long-data" or $n > p$ setting).

At the same time, the rise of national, continental, and intercontinental brain-data collections are making the statistical goal of prediction increasingly attractive. Modern machine-learning approaches have a focus on predicting disease status, behavior, even treatment response of single individuals. The process of deriving new knowledge based on a sample of participants takes a different form in the predictive analysis setting. Instead of looking within the sample of participants at the properties of the estimated parameters, the focus is on accurate statements about *new*, previously unseen participants—and evaluating the *out-of-sample generalization* (Vapnik 1998; Valiant 1984). In practice, the participants are split into two groups: a "training set" that is used to fit the model or classifier, and a separate "test set" that is used to evaluate prediction performance. If the prediction succeeds on the test set, we can empirically establish that the model captures useful biological structure and, more importantly, that a meaningful connection between (potentially many) input variables (e.g., fMRI brain scans) and a target variable (e.g., disease status) exists. Usually, the random split into train- and test-set is performed repeatedly in a procedure that is called *cross-validation*.

By quantifying the prediction success in new individuals (i.e., out-of-sample estimates) many machine learning approaches naturally adopt a prospective viewpoint and can directly yield a notion of clinical relevance. In contrast, classical approaches based on null-hypothesis testing often take a retrospective flavor as they usually revolve around finding statistical effects in the dataset at hand (so-called in-sample estimates) based on prespecified modeling assumptions, typically without explicitly evaluating some fitted models on unseen or future data points. Hence, ubiquitous techniques for out-of-sample generalization in machine learning are likely candidates for enabling a future of personalized psychiatry. This is because predictive models can be applied to and obtain answers from a single patient.

Two properties are shared between the discussed upcoming trends in data-analysis in the brain-imaging community. On the one hand, the anticipated shifts in statistical practice are expected to enable more complex (e.g., increased model expressiveness) and also more interpretable statistical models (e.g., more generative models) of the brain, based on high-dimensional neuroimaging data. On the other hand, many of these modeling approaches tend to work better with larger participant sample sizes and may be well prepared to handle rich high-dimensional input data. With the advent of the new data reality in the brain-imaging community, such "data-hungry" methods become increasingly feasible and necessary.

6.4 Clinical Endpoint Prediction in Single Psychiatric Patients Based on Brain-Imaging

In this last section, we place the trends of large-scale data collection and ensuing changes in statistical practice in the context of current mental health research. We give examples of how large-scale neuroimaging datasets can enable new research approaches and use a recent paper by Drysdale et al. (2017) to illustrate how parametric structure-discovery methods, latent factor models, and out-of-sample prediction all can be integrated in this type of research agenda.

The traditional approach to mental health research consists of identifying symptoms that frequently occur together and using these clinical manifestations to define disease-specific symptom combinations based on expert opinion. Clusters of symptoms are assumed to define coherent disease entities. These disease definitions are then used to find diagnostic biomarkers (e.g., by searching for neural correlates) or to predict treatment response. While this approach has worked well in many areas of medicine (consider, for example, the glomerular filtration rate to identify kidney disease) the same success has not yet materialized in psychiatry. Brain-based quantitative markers for predicting treatment response at the single-subject level, even to reliably distinguish between disease subtypes or healthy and diseased participants, remain elusive in mental health (Insel and Cuthbert 2015). Large-scale brain-imaging allows for flipping this approach on its head. Instead of clustering individuals into groups by clinical symptoms and then looking for neurophysiological correlates, we can cluster based on quantitative brain measurements directly (letting the brain data "speak for themselves") and then look at symptom measurements and clinical endpoints only after identifying clusters of shared brain dysfunction. As this alternative strategy underlies the ambition to directly model the biological basis of the disease and is less vulnerable to subjective and overlapping symptoms, it may be more likely to yield a reliable foundation for diagnosis and treatment.

Depression is one of many cases in psychiatry where recent evidence emphasizes unclear correspondence between diagnostic labels used in clinical practice and their neurobiological substrates as elucidated in neuroscientific research. Drysdale and colleagues employ functional neuroimaging to identify depression subtypes in brain biology (Fig. 6.4). In a large-scale study ($n = 1188$) they identified patterns of functional connectivity in resting-state fMRI that were associated with symptoms

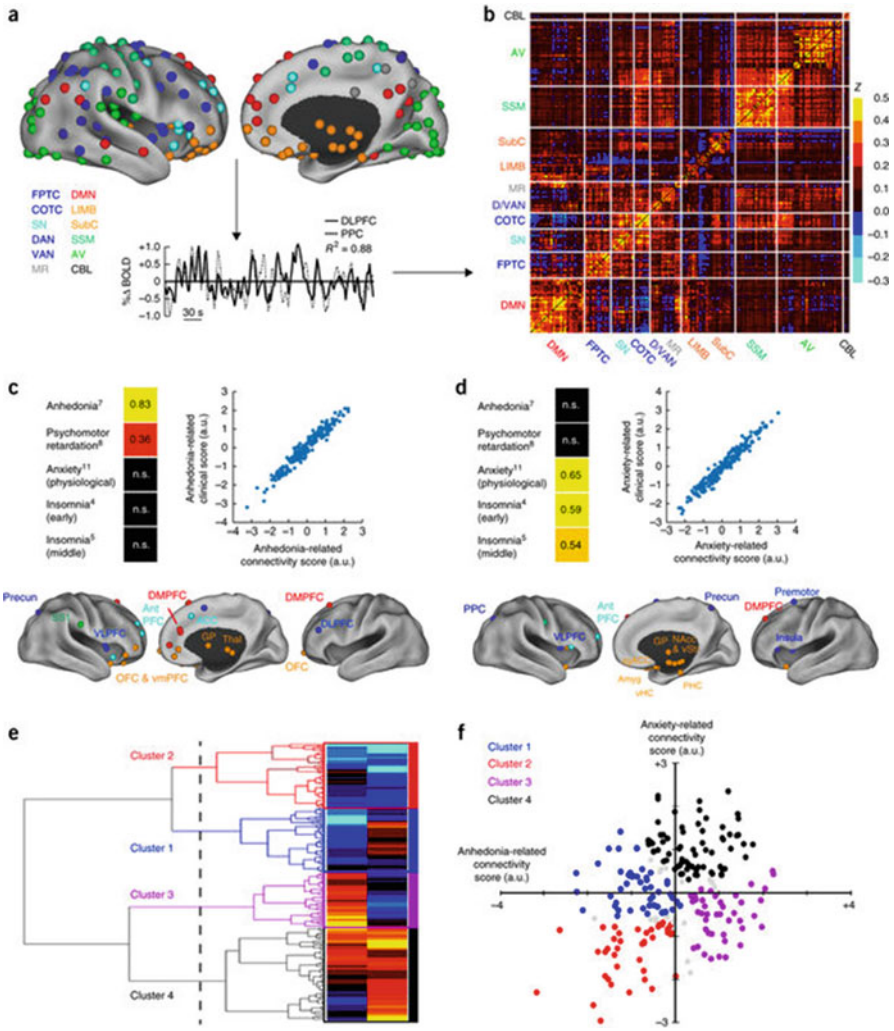

Fig. 6.4 *Example of modern brain-imaging-based subject stratification.* Neural activity time series measured by fMRI were extracted from regions of interest (**a**) and correlated with each other to yield "functional connectivity features" (**b**). Canonical correlation analysis was then used to find a small set of linear combinations of functional connectivity features that are maximally correlated with self-reported symptoms of depression (**c**, **d**). Thus, the number of variables per participant was reduced by two preparation steps: First from whole-brain maps to region-wise activity measures, then from functional connectivity features to even fewer components of variation obtained from CCA. This dimensionality reduction of high-resolution imaging data allowed identifying clusters of participants (**e**, **f**) which are predictive of distinct symptom-profiles and response to transcranial magnetic stimulation treatment. Reproduced from Drysdale (2017)

of depression and used these to identify four neurobiologically distinct subtypes of depression ("biotypes"). Based on these alternative group distinction for depressed patients they were then able to predict whether or not a patient would respond to transcranial magnetic stimulation (TMS)—a therapy in which a pulsing magnetic field is used to induce inhibitory or excitatory electric current into parts of the brain. The analysis approach in this study consisted of three steps: First, the authors built a *latent factor* model connecting fMRI and depression symptoms via CCA. Second, they used *parametric, discriminative* clustering to identify subtypes based on the previously derived latent factors. Third, they used support vector machines as a discriminative classifier to achieve *out-of-sample* predictions for the depression subtype based on fMRI data.

To better illustrate how the statistical methods tie into the quest for depression biomarkers we will cover the analysis pipeline more comprehensively. After preprocessing (the cortex and subcortical structures were parcellated into 258 regions of interest), resting-state fMRI time series were extracted for each region and correlated against each other. The resulting correlation coefficients (functional connectivity features) for each patient represented the left-hand side of the variable set for a canonical correlation analysis. The right-hand side of the variable set was given by the corresponding Hamilton Depression Rating Scale results for each patient. The CCA then returned hidden dimensions of variation—sets of distinct functional connectivity patterns correlated with distinct combinations of clinical symptoms. The number of latent factors was much smaller than the number of original regions, making the latent modeling results easier to analyze and interpret. The latent variability components were then used for clustering via the parametric *k*-means algorithm. This procedure used the similarity in functional connectivity to partition participants into k group such that each participant belonged to the cluster with the smallest mean distance. A split into four clusters appeared to provide useful partitioning solutions for defining maximally dissimilar patient subtypes.

Each of these subtypes (i.e., clusters derived from the latent factors) was shown to be correlated both with distinct patterns of abnormal functional connectivity as well as distinct clinical-symptom profiles. All four subtypes also featured shared functional connectivity patterns that corresponded to "core" symptoms that were present in all patients diagnosed with depression. The individual subtype predicted whether or not a given patient would respond to transcranial magnetic stimulation therapy. Support vector machines were trained to directly predict a patient's brain-derived subtype based on their functional connectivity information.

The steps of the analysis pipeline (latent factor model, clustering, prediction) were conducted on a training data set consisting of only two-thirds of the patients, in order to be able to test how well the discovered brain-behavior effect is likely to generalize to previously "untouched" data (the remaining one-third). That is, the built support vector machines prediction models were validated on the previously held-out test set and achieved accuracy rates of approximately 90% in predicting the biological subtype of individual patients—and thereby their individual response to TMS treatment. This study is one of the first proofs of concept that data-derived brain phenotypes of psychiatric disorders can provide useful biological categories that enable improved treatment choices on a single-subject basis.

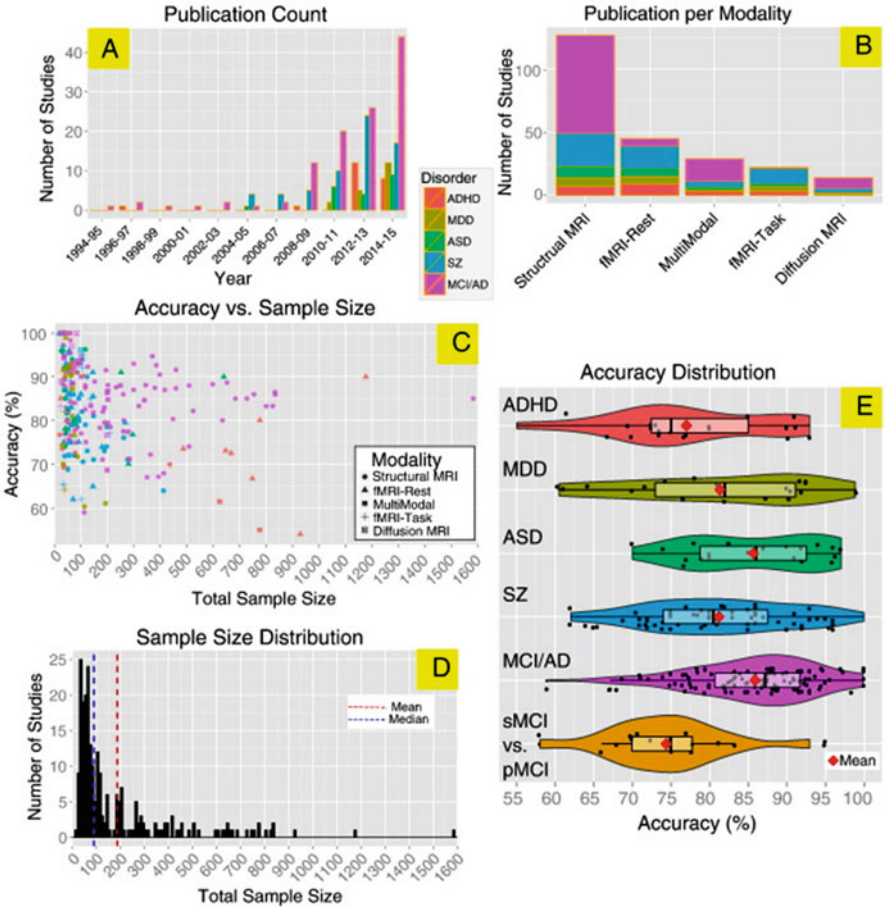

Fig. 6.5 *Single-subject prediction of brain disorders using neuroimaging.* A survey by Arbabshirani et al. (2017) shows strong growth in the number of brain-scanning studies that attempt to automatically classify brain disorders based on neuroimaging data (**a**). Structural MRI is so far the most frequently used input data for the purpose of classification (**b**). The number of participants is still relatively small (<200) for most imaging-based classification studies (**c**, **d**). Based on selected brain-imaging modalities and feature variables, different studies report diverging classification performances (**e**). Reproduced from Arbabshirani et al. (2017)

Over the last years, there has been a rising number of investigations into single-subject prediction of brain disorders in neuroimaging. Arbabshirani et al. (2017) recently provided a survey (Fig. 6.5) of ~200 recent studies. Based on their broad field analysis, structural and resting-state MRI are the brain-imaging modalities that are currently favored for predicting brain disorders, and most important brain disorders have been studied for single-subject prediction. Likely because of its severity and prevalence, mild cognitive impairment and Alzheimer's disease (MCD/AD) is the disorder that has most often been tried to predict based on

MRI data. The average prediction accuracy across studies was ~86% for MCD/AD and thereby yielded the comparatively best prediction accuracy among common brain disorders. Autism spectrum disorder yielded similar accuracies (~85%), followed by major depressive disorder and schizophrenia (~81%), and attention deficit disorder (~77%). Models in these studies were trained on relatively few participants (mean 186, median 88). Virtually all of these investigations had to restrict themselves to a correspondingly small number of features, usually derived by averaging brain regions via a brain atlas, or other biologically inspired manually crafted features. The reported average participant numbers were still an order of magnitude away from the projected number of (e.g., Alzheimer's) patients in the prospective UKBB study, leading us to anticipate further improvements in predictive accuracy and potential clinical applicability in diagnosis and prognosis of brain disorders as these data become available.

An intensified approach to psychiatric research based on brain-derived markers has several advantages over the traditional symptom-based research stream. Neuroimaging biomarkers can more directly allow gaining traction on neurophysiological aberrations underlying psychopathology. Identified brain-derived markers often enable reliable brain-based stratification of individual participants, which should offer a promising basis to improve clinical practice in diagnosis, prognosis, and treatment selection. Potential for more complete detection and exploitation of the pathophysiological mechanisms underlying brain disorders may fuel the development of new and superior treatment strategies. These anticipated advances may likely turn out to be a direct result of large-scale neuroimaging data collection combined with the use of data-hungry computational methods.

6.5 Conclusions

The soaring cost of psychiatric disease prompts a global urgency for finding new solutions (Bloom et al. 2012; Gustavsson et al. 2011). We believe that whether or not personalized medicine can be realized in psychiatry is largely a statistical question at its heart. For many decades, *the group* has served as the working unit of psychiatric research. Facilitated and intensified acquisition of always more detailed and diverse information on psychiatric patients is now bringing another working unit within reach—*the single patient*. Rather than pre-assuming disease categories and formally verifying prespecified neurobiological hypotheses, an increasingly attractive alternative goal is to let the data speak for themselves. As a consequence of the new data reality and changing research questions, some long trusted statistical methods may no longer be the best tool at our disposal.

The statistical properties of learning-algorithm approaches tailored for the data-rich setting promise clinical translation of empirically justified single-patient prediction in a fast, cost-effective, and pragmatic manner. Patient-level predictive analytics might also help psychiatry to move from strong reliance on symptom phenomenology to catch up with the biology-centered decision making in other branches of medicine. Machine learning tools offer an ideal data-guided framework

to uncover, foster, and leverage inter-individual variation in behavior, brain, and genetics. The fact that the currently embraced mechanistic explanations for psychiatric disorders range from molecular histone-tail methylation in the cell nucleus to urbanization trends in society as a whole highlights human-independent learning algorithms as an underexploited avenue for the automatic identification of disease-specific neurobiological features that can predict clinical outcomes. Ultimately, the human intelligence alone may be insufficient to decipher how mental disorders arise at the complex interplay between each individual's unique genetic endowment and world experience.

References

Alfaro-Almagro F et al (2018) Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage 166:400–424

Arbabshirani MR et al (2017) Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145:137–165

Bellman R (1957) Dynamic programming. Princeton University Press, Princeton

Berkson J (1938) Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc 33(203):526–536

Bickel PJ, Doksum KA (2007) Mathematical statistics: basic ideas and selected topics. Pearson, Upper Saddle River

Bishop CM (2006) Machine learning and pattern recognition. Information science and statistics. Springer, Heidelberg

Bishop CM, Lasserre J (2007) Generative or discriminative? Getting the best of both worlds. Bayesian Stat 8(3):3–24

Bloom DE et al (2012) The global economic burden of noncommunicable diseases. No. 8712. Program on the global demography of aging

Brodersen KH et al (2011) Generative embedding for model-based classification of fMRI data. PLoS Comput Biol 7(6):e1002079

Bzdok D, Meyer-Lindenberg A (2018) Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 3(3):223–230

Bzdok D, Yeo BTT (2017) Inference in the age of big data: future perspectives on neuroscience. NeuroImage 155:549–564

Bzdok D, Eickenberg M, Varoquaux G, Thirion B (2017) Hierarchical region-network sparsity for high-dimensional inference in brain imaging. Inf Process Med Imaging 10265:323–335

David O et al (2008) Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS Biol 6(12):2683–2697

Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, New York

Drysdale AT et al (2017) Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23(1):28–38

Editorial (2016) Daunting data. Nature 539:467–468

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

Efron B, Hastie T (2016) Computer age statistical inference. Cambridge University Press, Cambridge

Fisher RA, Mackenzie WA (1923) Studies in crop variation. II. The manurial response of different potato varieties. J Agric Sci 13(3):311–320

Focke NK et al (2011) Multi-site voxel-based morphometry—not quite there yet. NeuroImage 56(3):1164–1170

Freedman D (1995) Some issues in the foundation of statistics. Found Sci 1(1):19–39
Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning. Springer Series in Statistics, New York
Friston K, Penny W (2003) Posterior probability maps and SPMs. NeuroImage 19(3):1240–1249
Friston KJ, Penny W, Phillips C, Kiebel S, Hinton G, Ashburner J (2002) Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16(2):465–483
Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. NeuroImage 19(4):1273–1302
Gennatas ED et al (2017) Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to young adulthood. J Neurosci 37(20):5065–5073
Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
Gustavsson A et al (2011) Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 21(10):718–779
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman & Hall, London
Insel TR, Cuthbert BN (2015) Medicine. Brain disorders? Precisely. Science 348(6234):499–500
James G et al (2013) An introduction to statistical learning: with applications in R. Springer, New York
Jebara T (2012) Machine learning: discriminative and generative. Springer Science & Business Media, Berlin
Jordan MI (2011) A message from the president: the era of big data. ISBA Bull 18(2):1–3
Jordan MI et al (1999) An introduction to variational methods for graphical models. Mach Learn 37(2):183–233
Mejia AF, Nebel MB, Shou H, Crainiceanu CM, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA (2015) Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators. NeuroImage 112:14–29
Miller RG (1981) Simultaneous statistical inference. Springer, Heidelberg
Miller KL et al (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19(11):1523–1536
Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc Lond A Math Phys Sci 231:289–337
Smith SM, Nichols TE (2018) Statistical challenges in "big data" human neuroimaging. Neuron 97(2):263–268
Sudlow C et al (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3):e1001779
Takao H, Hayashi N, Ohtomo K (2013) Effects of the use of multiple scanners and of scanner upgrade in longitudinal voxel-based morphometry studies. J Magn Reson Imaging 38(5):1283–1291
Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142
Vapnik V (1998) Statistical learning theory. Wiley, New York
Varoquaux G, Gramfort A, Poline J-B, Thirion B (2010) Brain covariance selection: better individual functional connectivity models using population prior. Advances in neural information processing systems, pp 2334–2342
Wang H-T et al (2018) Dimensions of experience: exploring the heterogeneity of the wandering mind. Psychol Sci 29(1):56–71
Woo C-W et al (2017) Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 20(3):365–377
Yang Y, Wainwright MJ, Jordan MI (2016) On the computational complexity of high-dimensional Bayesian variable selection. Ann Stat 44(6):2497–2532