Trends in Neurosciences

CellPress
REVIEWS

Opinion

# Exploration, Inference, and Prediction in Neuroscience and Biomedicine

Danilo Bzdok[1,2,3,*,@] and John P.A. Ioannidis[4,5,*]

**Recent decades have seen dramatic progress in brain research. These advances were often buttressed by probing single variables to make circumscribed discoveries, typically through null hypothesis significance testing. New ways for generating massive data fueled tension between the traditional methodology that is used to infer statistically relevant effects in carefully chosen variables, and pattern-learning algorithms that are used to identify predictive signatures by searching through abundant information. In this article we detail the antagonistic philosophies behind two quantitative approaches: certifying robust effects in understandable variables, and evaluating how accurately a built model can forecast future outcomes. We discourage choosing analytical tools via categories such as 'statistics' or 'machine learning'. Instead, to establish reproducible knowledge about the brain, we advocate prioritizing tools in view of the core motivation of each quantitative analysis: aiming towards mechanistic insight or optimizing predictive accuracy.**

> '[Deep] neural networks are elaborate regression methods aimed solely at prediction, not estimation or explanation.' (Efron and Hastie [1], p. 371).

## The Emergence of Richer Datasets Alters Everyday Data-Analysis Practices

There is a burgeoning controversy in neuroscience on what data analysis should be about. Similarly to other biomedical disciplines, there are differing perspectives among researchers, clinicians, and regulators about the best approaches to make sense of the unprecedented data resources. Traditional statistical approaches, such as null hypothesis significance testing, were introduced in a time of data scarcity and have been revisited, revised, or even urged to be abandoned. Currently, a growing literature advertises predictive pattern-learning algorithms that are hailed to provide some traction on the data deluge [2,3]. Such tools for algorithmic predictions are increasingly discussed in particular fields of neuroscience ([4–9] for some excellent sources). Ensuing friction is aggravated by the incongruent historical trajectories of mainstream statistics and emerging pattern-learning algorithms – the former long centered on significance testing procedures to obtain *P* values, the latter with a stronger heritage in computer science [10–12]. We argue here that the endeavor of sorting each analytical tool into categories such as 'statistics' or 'machine-learning' is futile [13,14].

Take for instance ordinary linear regression, as routinely applied by many neuroscientists. The same tool and its underlying mathematical prosthetics can be used to achieve three diverging goals ([15], pp. 82–83; [16], chapter 4.12): (i) exploration, to obtain a first broad impression of the dependencies between a set of measured variables in the data at hand; (ii) inference, to discern which particular input variables contribute to the target variable beyond chance level; and (iii) prediction, to enable statements about how well target variables can be guessed based on data measured in the future.

## Highlights

As a prevalent misconception in neuroscience and biomedicine, null hypothesis significance testing is often thought to be the only existing, or most rigorous, framework for deriving reproducible conclusions from data.

Data analysis should be guided by the actual modeling goal. Exploration provides a first cursory glance that summarizes what can potentially be interesting in the data at hand. Inference typically focuses on isolating variables deemed individually important above some chance level, often based on *P* values. Prediction commonly aims at identifying variable sets that together enable accurate guessing of outcomes based on other or future data.

*P* values do not measure the predictive accuracy of a model. Variables declared important by null hypothesis significance testing can be incongruent with the variables that maximize predictive performance in new individuals or settings.

[1]Department of Psychiatry, Psychotherapy, and Psychosomatics, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, 52072 Aachen, Germany
[2]Jülich Aachen Research Alliance (JARA), Translational Brain Medicine, Aachen, Germany
[3]Parietal Team, Institut National de Recherche en Informatique et en Automatique (INRIA), Neurospin, Commissariat à l'Energie Atomique (CEA) Saclay, 91191 Gif-sur-Yvette, France
[4]Meta-Research Innovation Center at Stanford, Stanford University, Stanford, CA, USA
[5]Departments of Medicine, of Health

Confusion can arise because it is the motivation for using linear regression that differs between these scenarios. The mathematical mechanics underlying model parameter fitting are indistinguishable. Taken more broadly, instead of attaching labels of opposing camps to each analytical tool, it would be more productive, we would argue, to focus on the desired goal of a specific quantitative analysis. The goal, rather than the choice of a particular tool, is the major factor that ultimately determines what statements can confidently be made about brains, behavior, or genes, or, for that matter, any other question of interest.

## Exploration, Inference, Prediction: A Typology of Different Modeling Goals

The initial description of correlative relations in brain data is a common first step in many research projects. A crucial distinction arises when deciding on how to venture into identifying reproducible findings in quantitative analysis. How a particular analytical tool is used in a specific application domain may often be more important than which class of tool is chosen.

### Exploration of Correlative Associations

In various studies, a straightforward approach to charting candidate associations in brain data is Pearson's correlation (without computing *P* values). A simple statistic is thus computed between two series of measurements for descriptive purposes. As one concrete example, this analysis can quantify the relationship between amygdala activity measured in an fMRI experiment and some behavioral response. Such tentative data exploration can also be done in situations involving one input and one output variable by fitting a linear regression to the data. In these informal settings, the modeling goal is limited to a descriptive, correlational summary of the raw data that one happened to observe. Estimating linear-regression parameters alone does not license the importance of particular variable relationships (i.e., inference). Neither does a fitted linear regression itself declare whether these variable relationships hold up for other individuals or future datapoints (i.e., prediction).

### Inference of Statistically Significant (and Possibly Causal) Associations

Another goal is to try to isolate the specific contributions of single variables so as to reveal how the observed response depends on each particular measurement. This is a common agenda in many well-controlled experimental designs. For instance, in studies looking into the effects of a gene knockout in mice, or in clinical trials examining the impact of a specific treatment in patients. Historically, this type of deductive reasoning has often drawn on null hypothesis significance testing (NHST). The framework however is sometimes ill-suited and frequently misunderstood [17–19]. As an alternative to NHST, one may draw formal inference by means of false discovery rate (FDR), Bayesian posterior inference, or other tools ([1], chapters 3 and 15). Inferences also need to take into account various biases [20] to avoid making claims that represent false positives (in the NHST framework), underestimated FDR (in the FDR framework), or exaggerated posterior parameter distributions (in the Bayesian framework) ([1], chapter 3; [21], chapter 18.7). Much debate has emerged about what inferential statements about relevant variable contributions mean [10,13,22], and how significant associations tends towards the holy grail of uncovering causal influences [23].

### Generalization of Predictive Associations

One way to substantiate the explored correlations or inferred significance statements is by verifying whether these quantitative relationships still hold up for other datapoints or in new individuals. This goal is common to many observational, naturalistic, and prospective epidemiological studies. For instance, increasingly, predictive pattern-learning algorithms are used to derive the behavioral response of individuals from whole-brain neural activity or derive health risk from genomic profiling (*cf* [24–26]). Predictive modeling can also be carried out based on

Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University, Stanford, CA, USA
@Twitter: @danilobzdok

*Correspondence:
danilo.bzdok@rwth-aachen.de (D. Bzdok) and jioannid@stanford.edu (John P.A. Ioannidis).

standard linear regression. Several fields of clinical medicine have already accumulated a large literature of predictive scores and tools [27,28]. Currently, usage of predictive approaches lacks standardization and few extracted prediction rules are rigorously validated [29]. Even fewer are evaluated for replication in different settings and groups of individuals [30]. Increasingly complex predictive models use hundreds and thousands of parameters and/or try to benefit from non-linear interactions in extensive data such as electronic health records [31]. Notably, it has so far rarely been shown that accounting for complex non-linearity in 'big' medical data has considerably improved predictive performance. The low success rate is perhaps partly due to the still insufficient sample sizes or to limited quality of the measurements [7,20,32].

To be clear, exploration, inference, and prediction are not strictly mutually exclusive. Instead, quantitative investigations often involve a combination of the three approaches, prioritized to different degrees. In many neuroscience domains that are starting to amass 'big data', predictive pattern-learning algorithms are becoming popular alternatives to classic linear-regression applications [2,3]. Such algorithmic tools include support-vector machines, random forests, or artificial neural networks. Regardless of whether linear-regression approaches or pattern-learning algorithms are used, the main goal of the prediction enterprise is to put the built model, with already estimated model parameters, to the test against some independent data ([21], chapter 7). In this analysis regime, the investigator wishes to achieve the highest-possible forecasting performance. She is not necessarily worrying about how the model works or whether its fitted parameters carry biological insight.

## Inference and Prediction Serve Distinct Goals

Scientific insight has been a primary focus of the statistical methodology traditionally used in fields such as psychology and experimental neuroscience, as well as in assessments in evidence-based medicine. The underlying inferential approach is particularly well-suited for asking questions such as: which specific gene location contributes to or has an effect on a behavioral trait? Somewhat counterintuitively, in many cases genetic variants identified via such an inferential approach may not serve best to detect whether somebody has that behavioral trait or not [33,34]. This is because modeling for prediction typically asks a more heuristic type of question: which gene locations are collectively useful to distinguish individuals with or without the behavioral trait? Finding answers to this latter type of question follows the perhaps more superficial agenda of prioritizing successful recognition of any data relationships that are able to derive the specified outcome in independent individuals. Such predictive approaches put less emphasis on mechanistic insight into the biological underpinnings of the coherent behavioral phenotype (Table 1).

Inferring new scientific insight is often about answering questions such as: which input variable within a given dataset is an important contributor to the outcome? (or, is it a relatively more important contributor compared to other input variables?) Ideally, this modeling regime aims at mechanistic understanding of the impact of the input on the target variable. The investigator is interested in understanding the way in which an outcome $y$ is affected by a change in the input variables $x_1, \ldots, x_p$. To put it more mathematically, with $X$ denoting the measurement vectors $x_1, \ldots, x_p$, she wants to know 'how $y$ changes as a function of $x$' ([35], p. 19]). Consequently, inferential data analysis becomes difficult to the extent that the statistical model is a black box. Further, inferential statements about individual measurements of brain phenomena have their best chance of being reproducible if derived in the context of careful experimental controls (e.g., randomized trials in clinical assessments). Importantly, however, many, if not most, questions in neuroscience and biomedicine cannot even be addressed using randomization (*cf* [1], epilogue).

Table 1. The Inference–Prediction Continuum of Modeling Goals (*cf* Figure 1)

| Inference <-——————————————————————> Prediction | |
|---|---|
| Commonly Used Tools for Inference Goals<br>Null hypothesis significance testing to compute *P* values for specific target variables. Tools for this purpose include, for example, ANOVA, the *t* test, or $\chi^2$ test. Increasingly popular alternatives include false discovery rate and Bayesian posterior inference, as well as some pattern-learning algorithms (e.g., feature importance scores from random-forest algorithms). | Commonly Used Tools for Prediction Goals<br>Empirical validation schemes to compute prediction accuracy of the built model as a whole. Exemplary tools include support-vector machines, random-forest algorithms, and other ensemble and boosting techniques, the rapidly evolving 'deep'-learning algorithms, as well as ordinary and penalized linear regression. |
| Knowledge-Guided<br>Candidate variables are often hand-picked by the investigator in a targeted fashion based on existing substantive knowledge. Research questions are explicitly articulated before data collection in a carefully controlled experiment. The chosen variables are evaluated by an often simple but inflexible model that ideally is prespecified by the investigator before seeing the data. However, data dredging, and thus a high false-positive rate, are common in practice. | Pattern-Guided<br>A large and diverse array of 'found' variables is typically considered in the statistical analysis in a heuristic data-led fashion. It can be unknown how the data were generated, and the exact research question may be detailed as the data are being analyzed. The adaptive and sometimes very flexible model extracts a general prediction rule directly from the data in the spirit of 'letting the data speak for themselves'. |
| Explainable Narrative<br>Statements about the specific contribution of individual input variables are the priority. Such claims of variable relevance are often more readily available in simple linear-regression models. Accordingly, these models tend to be preferred in the context of inference such that every single parameter, and its corresponding unit, can be cleanly attributed its share of the explained variance. Usually, the meaning of each parameter should be readily understood, and hence the model often allows for a simplified narrative; statements are centered on single parameters rather than on the prediction performance of the collective model parameters. | Opaque Black Box<br>Although simple linear-regression models may perform reasonably well in terms of predictive power, if the goal is to maximize prediction accuracy, it is often beneficial to exploit complex non-additive associations in the data. In many real-world situations the target variable depends on the input variables in convoluted ways, which can hinder assigning to single input variables a clear relative contribution to the output, and model parameters are often treated as instrumental intermediates to achieve high prediction performance without necessarily aiming to assign specific meaning to each parameter estimate *per se*. |
| Formally Justified<br>Many traditional analysis techniques were rigorously characterized and validated by mathematical theory; simple linear models lend themselves well to theoretical model criticism, and carry well-understood modeling limits; another benefit of formal performance guarantees is the typically lower computational load. | Empirically Justified<br>Predictive models can be explicitly and quantitatively evaluated by applying the entire set of estimated model parameters to unseen independent, newly generated, or future observations or individuals; formal performance guarantees are often challenging; these models are often informally validated by means of more computationally demanding cross-validation, bootstrapping, and other resampling schemes. |
| Data-Efficient<br>Many methods from classical statistics were designed long ago to handle data that are scarce, as well as being laborious and expensive to collect. | Data-Hungry<br>Compared to classical statistics methods, many sophisticated predictive approaches require more data, especially when complex non-linear relationships are to be modeled and more hyperparameters need to be tuned; comparably more data also tend to be needed if each observation has many input variables, and if random noise is expected to be prominent (e.g., medical data). |
| Problem-Tailored<br>Each approach is designed to solve a particular data-analysis question, typically based on problem-specificprobabilistic and distributional assumptions about how the investigator believes the data have come about. | Versatile<br>Approaches are devised to provide useful solutions to various types of data and data-analysis questions. |

Quantitative analyses that strive to mechanistically explain the inner workings of brain phenomena have a different epistemic value than research aimed to model brain phenomena for the goal of accurate future predictions. In the prediction case, the investigator wants to extract knowledge about regularities by sieving through configurations of candidate patterns (and possibly very complex ones) [2,3]. Prediction accuracy is the core metric to capture how well the overall quantitative model – that is, the collection of fitted parameters – can emulate a high-level abstraction of mechanisms in nature. The predictive approach thus asks: how well can the built model reproduce the studied brain phenomenon that has been quantitatively captured in the measurements?

The priority to maximize prediction performance may require exploitation of more complex non-linear relationships in brain data, in contrast to widely adopted linear modeling. Recognizing complex relationships between variables is something that many black box pattern-learning algorithms excel at. The more transparent linear-regression approaches have served well in neuroscience and medicine, and are arguably epitomized in the successful era of genome-wide association studies (GWAS) [36]. By contrast, the data-led identification of predictive principles from non-linear relationships between variables has a strong legacy in the machine-learning community ([10]; [37], chapter 1.2).

A contrast between modeling goals lies in the readiness of non-linear predictive models to capture and capitalize on higher-order interactions among variables. Complex variable–variable (–variable– . . . ) interactions are probably common in brain phenomena. However, to best 'see' these higher-order interactions, the data need to be measured with little noise. When adequate data are available, more sophisticated analytical tools are generally advantageous in cases of higher-order variable interactions. Some non-transparent pattern-learning algorithms, capitalizing on non-linear interactions, have frequently ranked among the top solutions in international data-analysis competitions involving a diversity of challenging data types (e.g., www.kaggle.com). Many areas of brain research are experiencing advances in the granularity of quantitative measurements, increasing the potential to capture higher-order intervariable interactions. Thus, advanced pattern-learning algorithms may eventually outperform linear models even more often than is currently the case. Note, however, that the superiority of modeling complex patterns over simple linear approaches should not be taken for granted, and merits case-by-case evaluation. Altogether, compared to modeling for inference, the predictive analyst may favor tools that extract regularities from data in a way that is advantageous for prediction accuracy. High forecasting accuracy is favored even if opaque to human intuition, with 'deep' neural-network algorithms offering an extreme example of such tools.

Besides challenges added to parameter interpretation, predictive tools are typically less suited to detect causal relationships in data [23]. Nevertheless, a useful predictive model with high accuracy may be built based on measurements that are expected to have little causal relation to the outcome of interest. For instance, it has been acknowledged that 'Neuroimaging studies *per se* [ . . . ] only provide insights into neural correlates but not into neural causes of cognition' [38]. Neuroimaging measurements such as fMRI are only indirectly related to the dynamic activity changes in neuronal assemblies underlying cognitive processes. However, such signals carry intermediate information that can be used for accurate predictions of interindividual differences in cognition, such as propensity to attentional lapses, general intelligence, or health status [39,40].

To recapitulate, we have emphasized two distinct motivations that could drive a specific scientific inquiry: 'providing insight', for the purpose of inference, and 'accurately modeling

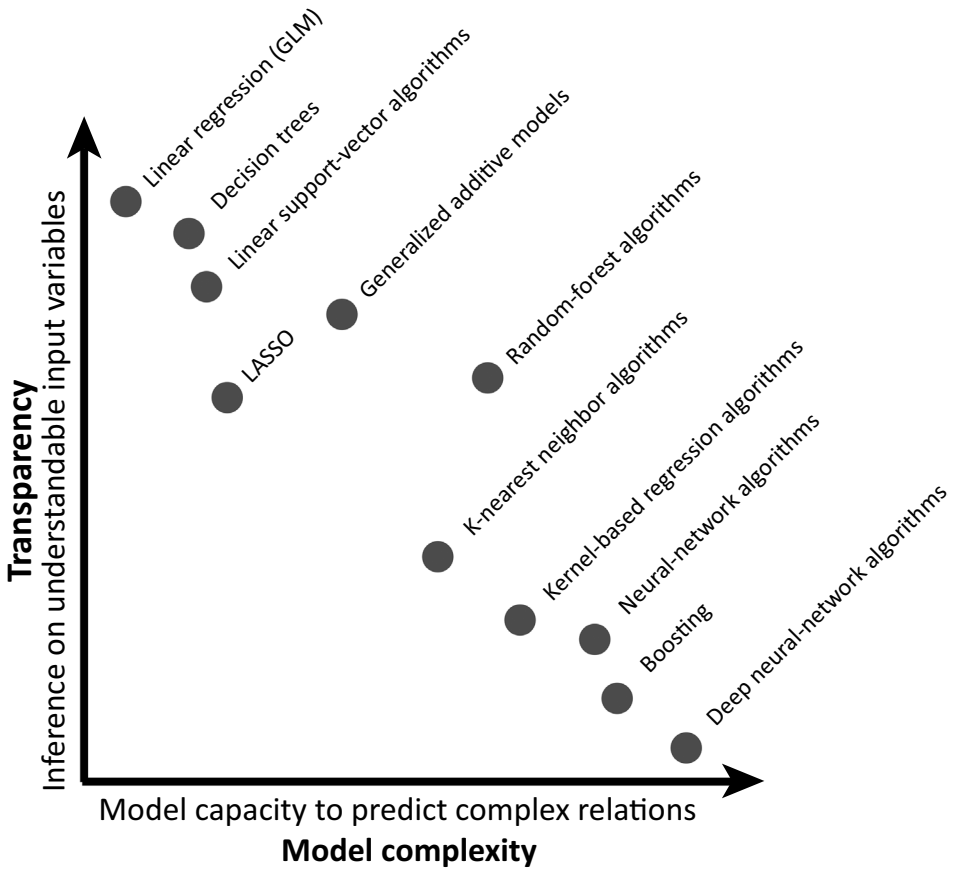


Figure 1. The Trade-Off between Model Transparency, Which Allows Scientific Understanding, and Theoretical Model Capacity, Which Affords Sophisticated Predictions. Neuroscience and biomedicine have had a long-dominating focus on scientific insight by using simple and thus transparent models. Such approaches are well suited to work towards the goal of inference regarding mechanistic understanding. This goal is epistemologically distinct from, and sometimes practically incompatible with, maximizing predictive power. The pragmatic goal of optimizing predictive accuracy can exploit large datasets even at the cost of opting for black box models that cannot easily be interrogated. In practice, the actual ratio between transparency and predictability depends on the specific analytical tool being used and the particular dataset at hand. Abbreviations: GLM, generalized linear models; LASSO, least absolute shrinkage and selection operator: a recently introduced constrained regression for high-dimensional data analysis, which is a special instance of GLM.

the world', for prediction. The inferential regime prioritizes statements about the relevance of each individual input variable. The predictive regime instead prioritizes the relevance of the output of the model (!) for precise forecasting. Predictive modeling describes what 'does' happen. Prediction often does not equally well address the question of 'how', and may be less apt for the question of 'why'. In addition, prediction is not always feasible and may remain mediocre in some applications, despite recent technical advances in data analytics. These considerations encourage trade-offs between model transparency for easy interpretability and model complexity that would enable predicting particularly complex relationships (Figure 1). One could make the case that some brain phenomena are so complex that impenetrable predictive pattern-learning algorithms may be all neuroscientists can hope for (*cf* [22]). Moreover, accelerating data aggregation and the wider availability of computation power are

opening a 'shortcut' path to useful outcome predictions, circumventing the traditional milestone of mechanistic discovery as an essential step towards effective predictive capabilities.

## Implications for Clinical Brain Research

Many clinical studies in brain research set out to identify variables that are statistically significantly associated with a disease. This includes significant differences in specific brain regions, their neural activity or anatomical abnormality, connections between brain regions, gene variants, and more. Deviations in such measurements in patients, however, may not always be the best-possible choices for building successful predictive approaches ([20]; [41], p. 185). This is perhaps not too surprising given that some questions beg modeling for the inference goal. For instance: which particular demographic indicator, ethnic background, or clinical parameter is robustly associated with adverse reaction of patients to a drug? The context of predictive modeling begs a different question at the heart of the study, even when using the same statistical technique. For instance: how well can we know in advance the risk in a particular patient for an adverse reaction to that drug? Predictive modeling regimes, we would argue, provide a natural path towards clinical relevance by immediately acting on clinical endpoints [42]. In fact, an official report of the American Statistical Association (ASA) emphasized that 'Statistical significance is not equivalent to scientific, human, or economic significance. Smaller *p*-values do not necessarily imply the presence of larger or more important effects, and larger *p*-values do not imply a lack of importance or even lack of effect.' [17].

Modeling for inference and prediction are two different tasks. Increasing this awareness will probably foster new research directions. Centering on clinical endpoint predictions can complement the quest for identifying the biological causes of disease. Historically, in research on the neural and genetic basis of brain disease, a prevailing philosophy has been to progress in two consecutive steps: discovery of new pathophysiological mechanisms, which are then used as a stepping stone to designing new targeted treatments [43]. Nevertheless, one might argue, after >50 years of biological research on the brain aimed at inference, there are relatively few definitively established etiopathological pathways. Neither are there many reliable biomarkers for most mental disorders [44].

Even in the ideal case of brain diseases caused by a single gene with considerable penetrance – such as the 22q11.2 deletion linked to schizophrenia risk [45], and the expansion of CAG triplet repeats linked to Huntington's disease [46] medical doctors could be assisted by patient-tailored predictive approaches. All individuals with such a genetic variant carry an escalated risk of developing the disease. However, various interindividual differences can still arise, including the timing of symptom onset, the constellation of symptoms displayed, disease severity, clinical trajectory, and treatment response. These clinical scenarios illustrate the distinction between the pursuit of scientific insight and the wish to forecast patient-specific disease manifestations – aiming at elucidating disease-causing biological mechanisms or creating prognostic value with relevance for medical care. Without doubt, there are potentially immediate gains from the pragmatic intention to search for signatures in complex data that can be exploited to predict clinical endpoints. Such a research program does not conflict with or belittle the importance of the longer-term endeavor to understand the primary biology of brain diseases.

Predictive approaches are increasingly adopted, recommended, and even expected by policy-makers [47,48]. However, there are several requirements before they can be considered to be suitable for wide application in real-world clinical settings (Box 1). Beneficial conditions for successfully translating new predictive approaches into clinical practice include the following:

### Box 1. Stages of Translating Predictive Approaches in Brain Research into Practice

(i) Model Building

To fit the parameters of the chosen predictive model, one first needs empirical measurements from the brain systems of interest. One common preparatory analysis is to probe variable–variable relationships using pairwise correlation plots. Another is to estimate genetic relatedness between the participants using principal component analysis of their genomic profiles. In behavioral experiments in animals or humans, exploratory data summaries can identify collinearity in response times. Such collinearity in response times foreshadows hindered statements about the relevance of individual experimental conditions (i.e., inference), but hardly affects forecasting condition response latencies in new participants (i.e., prediction).

(ii) Internal Validation

These procedures guard against overly optimistic modeling performances. Internal validation procedures, unlike external procedures (point iii), do not require new and independent data and are based only on the original subject sample or dataset that was used during model building [65]. Cross-validation and bootstrapping are resampling schemes ([21], chapter 7) that can estimate metrics of model quality [47], such as expected prediction accuracy for future data, uncertainty of parameter estimates, and variability of prediction errors. Indeed, 'working scientists often find the most interesting aspect of the analysis in the lack of fit rather than the fit itself' ([16], p. 92). Nevertheless, interindividual variability may still be underappreciated by using such internal validations alone [24].

(iii) External Validation

For stronger validation, predictive associations identified from the original subject sample or dataset need to be ascertained in other individuals or in datasets measured later [60,64,66]. Successful application of a predictive model of disease risk, for instance, requires validation in different groups of individuals [24,29]. This step is important to combat reproducibility issues [67]. Currently,external model validations are not done as often as they should be [68]. However, it is important to comprehensively benchmark the value of each predictive approach for clinicians, policymakers, and clinical guidelines [69]. For instance, external validation may be performed in different geographical areas, time periods, and settings (e.g., secondary vs primary care). Generally, some authors have proposed that 'the most stringent external validation involves testing a final model developed in one country or setting on subjects in another country or setting at another time. This validation would test whether the data collection instrument was translated into another language properly, whether cultural differences make earlier findings nonapplicable, and whether secular trends have changed associations or base rates' ([16], chapter 5.3.1).

(iv) Generalizability and Transposability

When evaluating the predictions of a model on new individuals, the more different these individuals are from the original subject sample, the stronger the test for generalizability [59,65]. Prediction accuracies are typically lower than in preceding steps. For instance, our ability to predict the clinical utility of drugs tends to be hindered for particular groups of patients, including women, children, and the elderly. Common comorbidities are also frequently under-represented or intentionally excluded in clinical studies. Meta-analysis methods can be useful for summarizing and examining the predictive performance of a model across different scenarios. Large datasets from multiple studies and electronic health records or registry databases provide promising opportunities for examining the generalizability of predictive approaches [70].

To enhance reproducibility, accurate and complete reporting is imperative for studies applying predictive models. Such reporting is crucial for being able to critically appraise predictive models, to perform acid-test validations of them, to evaluate their impact, and ultimately to translate them into clinical practice [27,71].

(i) Input variables for the predictive approach should be unambiguously defined as well as measured in a straightforward and standardized way.
(ii) Prediction performance needs to be better than what can be achieved using existing clinical methods for diagnosis and monitoring.
(iii) Accurate predictions need to be carefully validated in diverse settings [49]. It is important to accommodate variability that results from contextual factors such as circadian rhythm, menstrual cycle, and periods of stress.
(iv) The predictive approach must also show reproducibility in different groups of individuals and different ethnicities that did not contribute to model building. By analogy to drug treatments, a candidate predictive model may be found, for instance, to work better in

males than females or to be less effective in the elderly. Drug treatments can also have adverse effects in individuals with specific genetic profiles (*cf* [7]).

(v) Predictive successes can only result in better patient management and clinical outcomes if effective interventions are available. In Alzheimer's disease, for instance, a major current effort is directed to improving disease prediction years before symptom onset. Translating such prediction to better clinical outcome, however, would depend on whether treatment interventions are available that can leverage diagnosis in a much earlier stage of the disease.

(vi) Successful predictive models that are easy to use and transparent are likely to be adopted more readily by the medical community. Health professionals will probably avoid complex modeling approaches that are more difficult to interpret, require extra training, or depend on hard-to-obtain information.

(vii) Randomized clinical trials may need to certify the utility and safety of a new predictive approach for patients [50,51]. This cornerstone of evidence-based medicine will most likely continue to bolster clinical guidelines in the 'big data' era.

Finally, we outline various obstacles in the journey towards establishing predictive approaches for clinical management and intervention:

(i) When using medical data, strong non-linear effects have seldom been explicitly modeled or reported [52]. Even if complex interactions exist between measured variables, they may be difficult to extract from present day datasets, particularly those of still limited sample sizes [20]. Consequently, simple and less data-hungry predictive approaches are likely to remain among the go-to choices in many clinical settings. Elaborate predictive pattern-learning algorithms often cannot yet be used to their full potential, let alone 'deep' neural-network algorithms (*cf* [53]).

(ii) It is often difficult to know the optimal sample size for a particular prediction-oriented clinical research program beforehand. This limitation stands in contrast to the availability of power calculations in classical statistics. Reasons include the unknown complexity of the aspired prediction function, the amount of relevant input variables, and noise in the data ([20,54]; [55], p. 124).

(iii) A small signal-to-noise ratio plagues various forms of medical data. Examples of noisy measurements include readouts of histone modifications in genomics and brain activity changes scanned using fMRI, electroencephalography (EEG), or magnetoencephalography (MEG). As a rule of thumb, the more complex the predictive model, the higher its susceptibility to random variation in the data. Hence, in noisy data, it is trickier for advanced pattern-learning algorithms to identify reproducible signatures among the measured variables.

(iv) Similarly, flexible predictive pattern-learning algorithms are more prone to overfitting idiosyncrasies in the data, such as batch effects in multi-site studies [56]. To guard against fragile patterns, the various 'bells and whistles' of many of the sophisticated predictive approaches need to be chosen in a principled fashion [52]. These considerations invigorate the need for reproducible modeling practices as a core activity in computational biomedical research (*cf* [57]).

(v) The lack of transparency of predictive approaches that go beyond mainstream linear modeling is a particular concern that can erode the trust needed for implementation in clinical practice [47,52]. Indeed, skewed or wrong predictive approaches can systematically inflict harm by driving poor decision making [58].

(vi) Because of methodological constraints, much clinical brain research may not directly target real-world settings. Instead, clinical studies routinely enroll patients based on stringent exclusion criteria such as medication use or common comorbidities. These study designs may impede our ability to make predictions in realistic clinical settings. For instance, assessing the effectiveness of drugs or other treatments is particularly hindered when it comes to patient groups that are relatively rarely recruited in clinical studies, such as children and the elderly [59].

(vii) Electronic health records are soon likely to provide rich resources to build effective predictive approaches. However, there is still a scarcity of standardized health records involving large samples of patients. In addition, a bias towards sicker people has been noted in the few existing studies using such patient data being gathered by medical institutions [30,60].

## Concluding Remarks and Future Perspectives

The advent of 'big data' in neuroscience and biomedicine has started to transform many important sectors. In the 21st century, large-scale data aggregation, catalyzed by new modes of data dissemination and open science [61], has reached an unprecedented scale. Nonetheless, it remains unclear whether these emerging opportunities also prompt a deeper revision of the traditional 'value system' pertaining to scientific evidence. The data-rich neuroscientist can ask many new questions that could probably never be addressed quantitatively before. We encourage investigators and clinicians to rethink data analysis in the context of a repertoire of modeling goals (see Outstanding Questions). Choosing a data-analytic strategy for a research question at hand should not be a matter of tradition, habit, or taste.

It is worth reiterating that a specific analytical tool can serve multiple modeling goals. Linear regression, for instance, has been often used for exploratory summaries of possible relationships among measured variables. The same tool, however, can be used for inferring the most relevant mechanistic candidates among the measured variables, as well as for predicting outcomes by applying the built regression model to new datapoints. Conversely, many machine-learning algorithms have a long-standing track record in serving the predictive goal. Nevertheless, despite the increased complexity of many of these algorithmic tools, they can also be used towards the aim of data exploration, or even inference to isolate individually important input variables.

More broadly, as with any scientific method, modeling for either inference or prediction each comes with strengths and weaknesses [19,34,62,63]. Inferential modeling has been an established practice for decades [50,64]. By contrast, the most effective use cases still need to be identified for deploying predictive approaches in neuroscience and personalized medicine. Ultimately, deducing scientific insights and making pragmatic predictions are intimately related, but also importantly different.

### Outstanding Questions

Investigators in neuroscience and biomedicine would benefit from a principled understanding of how inference and prediction converge and diverge in everyday data analysis. When do statistically significant variables usefully contribute to accurate predictions? When are variables found to be predictive but not declared to be statistically significant? And when can variables serve both these different modeling goals?

Simple linear models are an optimal tool for the purpose of classical statistical inference. Complicated non-linear interactions in the data can potentially be extracted by more sophisticated predictive pattern-learning algorithms. In which types of biomedical data and in which application contexts can elaborate predictive approaches empirically outperform traditional linear models?

Predictive approaches can achieve high forecasting accuracy even if the variables measured have no obvious causal relation to the outcome and even though the extracted prediction rule may remain obscure to human understanding. How can we identify, discuss, and tackle the pressing ethical and legal consequences of predictively successful, but mechanism-naive, black box modeling such as using 'deep' neural-network algorithms?

### Acknowledgments

We are grateful to Jérémy Besnard-Lefort, Avram Holmes, Hannah Kiesow, Timm Poeppl, Marc-Andre Schulz, Bertrand Thirion, and Thomas Wiecki for insightful comments on a previous prevision of the manuscript. We thank three anonymous reviewers for many thought-provoking comments. D.B. is funded by the Deutsche Forschungsgemeinschaft (DFG; BZ2/2-1, BZ2/3-1, and BZ2/4-1; International Research Training GroupIRTG2150), Amazon AWS Research grants (2016 and 2017), as well as the START-Program of the Faculty of Medicine (126/16) and Exploratory Research Space (OPSF449), RWTH Aachen.

### References

1. Efron, B. and Hastie, T. (2016) *Computer-Age Statistical Inference,* Cambridge University Press
2. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
3. Jordan, M.I. and Mitchell, T.M. (2015) Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260
4. Haynes, J.-D. (2015) A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron* 87, 257–270
5. Kriegeskorte, N. and Douglas, P.K. (2018) Cognitive computational neuroscience. *Nat. Neurosci.* 21, 1148–1160
6. Marblestone, A.H. *et al.* (2016) Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* 10, 94
7. Woo, C.-W. *et al.* (2017) Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* 20, 365–377
8. Pereira, F. *et al.* (2009) Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, 199–209
9. Naselaris, T. *et al.* (2011) Encoding and decoding in fMRI. *Neuroimage* 56, 400–410
10. Breiman, L. (2001) Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–231

11. Donoho, D. (2017) 50 years of data science. *J. Comput. Graph. Stat.* 26, 745–766
12. Bzdok, D. (2017) Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.* 11, 543
13. Blei, D.M. and Smyth, P. (2017) Science and data science. *Proc. Natl. Acad. Sci.* 114, 8689–8692
14. Jordan, M.I. *et al.* (2013) *Frontiers in massive data analysis,* The National Academies Press
15. Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models (Monographs on Statistics and Applied Probability Vol 43),* Chapman & Hall
16. Harrell, F.E. (2001) *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression,* Springer
17. Wasserstein, R.L. and Lazar, N.A. (2016) The ASA's statement on p-values: context, process, and purpose. *Am. Stat.* 70, 129–133
18. Szucs, D. and Ioannidis, J. (2017) When null hypothesis significance testing is unsuitable for research: a reassessment. *Front. Hum. Neurosci.* 11, 390
19. Amrhein, V. *et al.* (2017) The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ* 5, e3544
20. Steyerberg, E.W. *et al.* (2018) Poor performance of clinical prediction models: the harm of commonly applied methods. *J. Clin. Epidemiol.* 98, 133–143
21. Hastie, T. *et al.* (2001) *The Elements of Statistical Learning,* Springer Series in Statistics
22. Norvig, P. *et al.* (2017) On Chomsky and the two cultures of statistical learning. In *Berechenbarkeit der Welt?* (Pietsch, W., ed.), pp. 61–83, Springer
23. Pearl, J. (2009) Causal inference in statistics: an overview. *Stat. Surv.* 3, 96–146
24. Fusar-Poli, P. *et al.* (2018) The science of prognosis in psychiatry: a review. *JAMA Psychiatry* 75, 1289–1297
25. Bzdok, D. and Yeo, B.T.T. (2017) Inference in the age of big data: future perspectives on neuroscience. *NeuroImage* 155, 549–564
26. Rosenberg, M.D. *et al.* (2018) Prediction complements explanation in understanding the developing brain. *Nat. Commun.* 9, 589
27. Siontis, G.C. *et al.* (2011) Predicting death: an empirical evaluation of predictive tools for mortality. *Arch. Intern. Med.* 171, 1721–1726
28. Siontis, G.C. *et al.* (2012) Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ* 344, e3318
29. Siontis, G.C. *et al.* (2015) External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J. Clin. Epidemiol.* 68, 25–34
30. Goldstein, B.A. *et al.* (2017) Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* 24, 198–208
31. Rajkomar, A. *et al.* (2018) Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* 1, 18
32. Arbabshirani, M.R. *et al.* (2017) Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage* 145, 137–165
33. Shmueli, G. (2010) To explain or to predict? *Stat. Sci.* 289–310
34. Lo, A. *et al.* (2015) Why significant variables aren't automatically good predictors. *Proc. Natl. Acad. Sci. U. S. A.* 112, 13892–13897
35. James, G. *et al.* (2013) *An Introduction to Statistical Learning,* Springer
36. Visscher, P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22
37. Goodfellow, I.J. *et al.* (2016) *Deep Learning,* MIT Press
38. Weichwald, S. *et al.* (2015) Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage* 110, 48–59
39. Gabrieli, J.D. *et al.* (2015) Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* 85, 11–26
40. Finn, E.S. *et al.* (2015) Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* 18, 1664–1671
41. McElreath, R. (2015) *Statistical Rethinking,* Chapman & Hall/CRC
42. Paulus, M.P. (2015) Pragmatism instead of mechanism: a call for impactful biological psychiatry. *JAMA Psychiatry* 72, 631–632
43. Insel, T.R. and Cuthbert, B.N. (2015) Brain disorders? Precisely. *Science* 348, 499–500
44. Weinberger, D.R. and Radulescu, E. (2015) Finding the elusive psychiatric 'lesion' with 21st-century neuroanatomy: a note of caution. *Am. J. Psychiatry* 173, 27–33
45. Bassett, A.S. and Chow, E.W. (2008) Schizophrenia and 22q11.2 deletion syndrome. *Curr. Psychiatry Rep.* 10, 148
46. Bates, G.P. *et al.* (2015) Huntington disease. *Nat. Rev. Dis. Primers* 1, 15005
47. Moons, K.G. *et al.* (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* 162, W1–W73
48. Bzdok, D. and Meyer-Lindenberg, A. (2018) Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 3, 223–230
49. Moons, K.G. *et al.* (2012) Risk prediction models. I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart* 98, 683–690
50. Ioannidis, J.P. and Tzoulaki, I. (2010) What makes a good predictor? The evidence applied to coronary artery calcium score. *JAMA* 303, 1646–1647
51. Paulus, M.P. *et al.* (2016) A roadmap for the development of applied computational psychiatry. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 1, 386–392
52. Steyerberg, E.W. *et al.* (2014) Risk prediction with machine learning and regression methods. *Biom. J.* 56, 601–606
53. He, T. *et al.* (2018) Is deep learning better than kernel regression for functional connectivity prediction of fluid intelligence? In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pp. 1–4, IEEE
54. van Smeden, M. *et al.* (2018) Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat. Methods Med. Res.* Published online January 1, 2019. http://dx.doi.org/10.1177/0962280218784726
55. Abu-Mostafa, Y.S. *et al.* (2012) *Learning from Data,* AMLBook
56. Steyerberg, E.W. and Vergouwe, Y. (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur. Heart J.* 35, 1925–1931
57. Poldrack, R.A. *et al.* (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* 18, 115
58. Shah, N.D. *et al.* (2018) Big data and predictive analytics: recalibrating expectations. *JAMA* 320, 27–28
59. Naci, H. and Ioannidis, J.P. (2015) How good is 'evidence' from clinical studies of drug effects and why might such evidence fail in the prediction of the clinical utility of drugs? *Annu. Rev. Pharmacol. Toxicol.* 55, 169–189
60. Djulbegovic, B. and Ioannidis, J.P. (2019) Precision medicine for individual patients should use population group averages and larger, not smaller, groups. *Eur. J. Clin. Invest.* 49, e13031
61. Leonelli, S. (2016) *Data-Centric Biology: A Philosophical Study,* University of Chicago Press
62. Wu, T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25, 714–721
63. Bzdok, D. *et al.* (2018) Statistics versus machine learning. *Nat. Methods* 15, 233–234
64. Manrai, A.K. *et al.* (2018) In the era of precision medicine and big data, who is normal? *JAMA* 319, 1981–1982

65. Steyerberg, E.W. and Harrell, F.E. (2016) Prediction models need appropriate internal, internal–external, and external validation. *J. Clin. Epidemiol.* 69, 245–247

66. Austin, P.C. *et al.* (2016) Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J. Clin. Epidemiol.* 79, 76–85

67. Nosek, B.A. *et al.* (2015) Promoting an open research culture. *Science* 348, 1422–1425

68. Studerus, E. *et al.* (2017) Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting. *Psychol. Med.* 47, 1163–1178

69. Damen, J.A. *et al.* (2016) Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 353, i2416

70. Riley, R.D. *et al.* (2016) External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 353, i3140

71. Iglesias, A.I. *et al.* (2014) Scientific reporting is suboptimal for aspects that characterize genetic risk prediction studies: a review of published articles based on the Genetic Risk Prediction Studies statement. *J. Clin. Epidemiol.* 67, 487–499