**On the prediction-inference dilemma in biomedicine**

Danilo Bzdok1,2,3,\* Denis Engemann3, Olivier Grisel3, Gaël Varoquaux3, Bertrand Thirion3

1 Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, 52072 Aachen, Germany  
2 JARA-BRAIN, Jülich-Aachen Research Alliance, Germany  
3 Parietal team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France

\* corresponding author: Prof. Danilo Bzdok, MD, PhD

Email: [danilo.bzdok@rwth-aachen.de](mailto:danilo.bzdok@rwth-aachen.de)

Phone: +49 241 80-85729

Universitätsklinikum Aachen

Pauwelsstr. 30

52074 Aachen

GERMANY

# Abstract

Many achievements of empirical research and evidence-based medicine in the 20th century were grounded in p-values and accompanying methods. In the 21st century, growing ambitions towards precision medicine put a premium on accurate predictions for single patients. This shift incurs tension between established statistical tools used to infer group effects and emerging machine-learning tools to achieve accurate predictions for individuals. Here, we provide an explicit comparison between linear regression for identifying significant contributing factors and learning algorithms for selecting predictive measures. In artificial data simulations and common medical datasets, we quantitatively characterized instances when inference and prediction concur and diverge. While both modeling approaches allowed for related conclusions, we describe divergence several data-analysis settings: certain variables turned out to be predictive but not significant, or significant but not predictive. More complete understanding of different ways to define importance is a prerequisite for biomedical research findings that are reproducible and exploitable for personalizing clinical care.

**Keywords**: scientific discovery | statistical significance | prediction performance | variable importance | learning algorithms | artificial intelligence

# Introduction

Inference and prediction can serve distinct purposes in the scientific inquiry of human health and disease ([1-3](#_ENREF_1)). Let’s take diabetes mellitus as an example. The inference paradigm can be used to establish biological facts that provide insight into the biological pathways that contribute to disturbed blood sugar levels (hyperglycemia). Diabetes can be a result of insufficient production of insulin hormone in the pancreas (type 1, onset mostly in children). Diabetes may also result from deficient insulin receptor response in body cells (type 2, onset mostly in adults). Diabetes can moreover affect previously healthy pregnant women (gestational diabetes). The clinical manifestation of disturbed blood glucose probably underlies partly diverging pathophysiology, which motivate other therapeutic interventions with statistically significant benefit. Type 1 diabetes can be treated by injecting missing insulin, while type 2 diabetes can be counteracted by surgery in obese patients. In turn, diabetes developed in the pregnant patient group usually resolves without treatment after delivery.

Instead of certifying the “trueness” of effects in disease and treatment, the prediction paradigm aims to detect statistical regularities that hold in the future. Diabetes can be automatically diagnosed based on frequent urination or increased thirst, possibly combined with age and gender, or some of the later consequences, including retina damage or kidney alterations. Recognizing symptom patterns is possible without understanding the biological processes that led to or maintain the disease. Further, a pattern-extraction algorithm may detect lacking production of insulin (type 1), or presence of pregnancy in women. However, the identified hints allowing reliable detection of diabetes type may shed limited light on the biological underpinnings. In treatment, an insulin pump could be engineered that achieves nuanced forecasting of sugar response regularities specific to the metabolism of a particular patient. Similar personalized profiling may enable risk prognosis and early intervention before onset of symptoms or long-term effects to improve medical care without understanding the biological pathways at play. In this way, both inference and prediction have important contributions to make to biomedical research - we want to know how a disease works and we want to know what will happen next.

Inference is intimately linked to statistical null-hypothesis testing and interpreting p-values. This data-analysis framework emerged in the early 20th century and is closely related to tools like linear regression, *t-*tests, and ANOVA. Electrical calculators not yet available ([4](#_ENREF_4), [5](#_ENREF_5)), this was a time when data were rare and expensive to acquire ([4](#_ENREF_4), [6](#_ENREF_6)). Research experiments were therefore carefully designed in advance and well-controlled. The historical context also explains why classical inference was intended to ask research questions in small samples that can be addressed by interpretable handpicked statistical models with few knobs to tweak (i.e., parameters) ([7](#_ENREF_7)). Many early statistical inventions were especially tuned to yield understanding of the relationship between a few candidate measures. Most medical doctors and biomedical researchers have been “raised” with this statistical culture during university training. If the goal is to examine whether an effect exists or which specific input variables have most impact on an output variable, classical null-hypothesis testing is still the gold standard today. Some investigators have however cast doubt that computing p-values to drawn statistical inference will play an invariably important role for biomedical research throughout the 21st century. John Ioannidis stated ([8](#_ENREF_8)): "With the advent of big data, statistical significance will increasingly mean very little because extremely low P values are routinely obtained for signals that are too small to be useful even if true."

Around the turn of the century, the rapidly increasing availability of whole-genome sequencing techniques and high-resolution body scanning ushered biology and medicine into the era of “big data”. There is a growing interest in and pressure for the creation, curation, and collaboration of extensive medical datasets. For instance, the UK Biobank has gathered genetic and environmental (e.g., nutrition, lifestyle, medications) data from 500,000 volunteers, and is the currently largest biomedical data resource of its kind. As a consequence of the recent co-occurrence of data availability, computing power, and cheaper data storage ([9](#_ENREF_9), [10](#_ENREF_10)), the realm of data-analysis has expanded more in the last two decades than probably ever before ([7](#_ENREF_7)). Flexible predictive algorithms have been specifically tailored for searching through massive data to extract subtle patterns ([6](#_ENREF_6)). Such predictive pattern-learning approaches promise improved clinical translation of empirically justified single-patient prediction in a fast, cost-effective, and pragmatic manner. Indeed, pioneering studies have leveraged predictive "deep learning" algorithms ([11](#_ENREF_11" \o "Hinton, 2006 #5956)) to i) estimate the cardiovascular risk, blood pressure, and smoking behavior from signs in retina scans using medical data from almost 300,000 patients ([12](#_ENREF_12" \o "Poplin, 2018 #7026)), ii) detect different heart arrhythmia as well as cardiologists in electrocardiograms from 30,000 patients ([13](#_ENREF_13" \o "Rajpurkar, 2017 #7027)), and iii) diagnose malignant skin cancer as well as dermatologists using almost 130,000 pictures ([14](#_ENREF_14" \o "Esteva, 2017 #6829)).

However, it is important to appreciate that the potential immediate gains of the pragmatic goal to identify patterns useful to predict clinical endpoints in complex medical data does not preclude the longer-term research agenda to understand the primary biology of diseases like diabetes. Carefully designed, meticulously conducted, and expansive experiments to confirm or reject a-priori verbalized research hypotheses in animals and humans will probably remain a cornerstone to generate biomedical knowledge.

**Methods**

**What do we mean by ‘inference’?**

As the term has been borrowed by various scientific fields to mean different things ([6](#_ENREF_6)), we want to make clear that we adopt the technical meaning common in classical null-hypothesis testing ([15](#_ENREF_15)). Inference is aimed at scientific discovery by uncovering certain “true” properties about a natural phenomenon of interest by answering whether an effect is likely to exists in the world. This modeling goal is for instance especially suited to ask, “Which gene locations *contribute to* or *are associated* with a disease?” Providing such novel insight as a service to science is achieved by making explicit assumptions about how the observed data arose. Properties of the underlying probabilistic model are then derived by understanding the way the outcome is affected by a set of measures of interest. The inference paradigm is aimed at better understanding the individual relevance of each input measure in impacting the response variable. In particular, the investigator wants to quantify the relatively more important predictors among a large set of hand-selected candidate variables. This intention explains why historically many statistical approach in the empirical sciences have been linear model approaches, even if the “true” relationship in nature may be more complicated. The modeling agenda is self-consistent in assuming that the quantitative model is a sufficient, fully specified summary of the phenomena under study. Often combined with careful experimental control and backed up by formal theory, modelling for inference is how traditional academic statistics have routinely dealt with small to medium datasets.

**What do we mean by ‘prediction’?**

Ascertaining properties of the inner workings of the studied phenomenon is importantly different from the prediction goal in empirical research. Here, the emphasis is on accurately modeling the world ([16](#_ENREF_16" \o "Hastie, 2001 #3957), [17](#_ENREF_17" \o "Jordan, 2015 #5958)). The investigator wants to automatically extract knowledge of regularities in the world searching through possibly meaningful patterns. This modeling goal is for instance especially suited to ask, “Which gene locations are *useful* to *distinguish* diseased versus healthy individuals?” Prediction accuracy is the core metrics to capture how well the quantitative model can emulate a high-level description of mechanisms in nature, that is, how well the model can reproduce the studied phenomenon whose data is analyzed. In the extreme case, the quantitative model may embody the discovered statistical relationship in a way that is opaque to the investigator. The prediction paradigm achieves guesses with high accuracy as those models are expected to generalize extracted patterns onto tomorrow’s data. There is smaller concern for what the achieved prediction means for the general population from which the sample was drawn. The quantitative model is used for prediction in new individuals whose outcome information we do not yet have. Typically, the predicted outcomes cannot be easily obtained, are expansive, or hard to come by. This aspect of “filling in” missing information also explains why mere correlation between two variables, such as in Pearson’s correlation, may be a more limited notion of foretelling future, yet-to-be measured observations ([18](#_ENREF_18" \o "Bzdok, 2018 #7022)). Prediction has been an important focus of activity in the more recent “machine-learning” community ([2](#_ENREF_2" \o "Breiman, 2001 #4148)) and corresponds to how data analysis is often practiced in data-intensive industry ([19](#_ENREF_19" \o "Henke, 2016 #6718)).

**Using the linear model for inference**

To assess which variables are statistically significant related to an outcome, we used multiple linear regression. This probably most common approach to perform least-squares regression optimized the following objective:

where is the number of individuals who contributed to the dataset, is the number of input variables (called *independent* or *explanatory variables*) measured for each individual, and is the outcome measure (called *dependent* or *explained variable*) that is to be expressed as a weighted sum of the variables . The data were standardized by mean centering to zero and variance scaling to one. This linear combination is estimated by fitting the (randomly initialized) coefficients to the observations in the dataset. The approach can answer questions about the relative contributions of each of the input variables in explaining the output y. Mechanisms in the data are assumed to be sufficiently described by means and variances as parts of the probability model ([15](#_ENREF_15)). The fitted model is assumed to encapsulate a complete description of how the particular input measures increase or decrease in parallel with each other to collectively explain variability in the response variable.

After model estimation, statistical inference was drawn as a second step to decide whether the contribution of input variable in explaining the response is sufficiently important to be *significant*. The relevance of the effects is computed based on the confidence intervals of the beta coefficients {Gelman, 2007 #7004}. Inferential conclusions are drawn by formally testing for the existence of an effect expressed under the null-hypothesis (e.g., a gene is not associated with schizophrenia) in opposition to the alternative hypothesis (e.g., a gene is associated with schizophrenia). The ensuing *p*-value indicates whether data from the subject sample at hand are too extreme to occur under the null hypothesis. Each of them corresponds to the null hypothesis that the beta at hand deviates from zero, whereas the other model coefficients do not. A non-significant beta coefficient suggest that the variable can be dropped from the model with little or no worse explanation. In typical applications of null-hypothesis testing, the p-value is computed on the entire data from *all* considered subjects.

**Using the linear model for prediction**

For comparison with ordinary linear regression, we chose the LASSO as a minor modification to turn it into a predictive pattern-learning algorithm {Tibshirani, 1996 #5961}. The linear models are almost the same, but the goal is different. Automatic variable selection is achieved by minimizing a very similar optimization objective:

where is the number of individuals who are included in the dataset, is the number of input variables (in this context often called *features*) measured for each individual, and is the outcome measure (called *target variable*) that is to be expressed as a weighted sum of the standardized variables . This linear combination is estimated by fitting the randomly initialized coefficients to the observations in the dataset. The hyper-parameter controls the amount of sparsity imposed on the model fitting. The higher the higher the tendency to set specific coefficients to exactly zero, which effectively “silences” the corresponding measures influence on explaining the output variable. A probability model is not “required” - whether the confidence intervals exceeded a threshold or not is here often no optimality criterion for variable importance. We also do not assume that means and variances full describe the probabilistic mechanissm in the data, only that they are informative enough to make useful predictions about the future. The confusion thing is that it is the motivation that is utterly different, the mathematics of the optimization objective is the same (if ), there is a key difference in perspective. Once fitted, the model can be applied to other samples to *predict* unobserved outputs or ”shipped” to other laboratories for repeated application. The chosen model automatically chooses the minimal subset of variables necessary for classifying for instance healthy versus diagnosed individuals. At its extreme, we do not use beta because we just use them as an intermediate step to achieve prediction, not because we care about this parameter itself so much.

Following model estimation, the importance of the candidate model is evaluated based on empirical guarantees. Answering the question whether an obtained predictive algorithms generalizes to unseen data is tackled in a heuristic fashion. It is typically achieved by identifying relationships in one set of subjects as a function of how these patterns persists in other individuals from a different set of subjects. Here, model parameters are typically estimated on some data while the emerging model is explicitly put to the test in some independent data from unseen individuals ([20](#_ENREF_20)). Explicit model checking was performed by evaluating the prediction performance of learning algorithms is typically performed by a procedure called *cross-validation* ([20](#_ENREF_20)). First, the machine-learning algorithm is built on a larger part of the dataset. Second, emerging candidate algorithms are evaluated and selected on unused data ([16](#_ENREF_16)). Because all conditions for independent, identically distributed observations are usually met for the left-out data, the out-of-sample prediction performance on the testing data samples can quantify how likely the same pattern could be detected in future, not yet seen patients. We care much more about a model's performance on the test data set than the training data set, since its performance on the test data set is much more likely to predict how the model will do on (other) unseen data. This approach to draw rigorous conclusion from data with the linear model assesses the robustness of patterns between typically many variables by testing how well an already fitted model extrapolates to unseen brain measurements. In practice, cross-validation procedures are frequently used to quantify out-of-sample performance by an unbiased estimate of a model's capacity to generalize to data samples acquired in the future.

This analysis paradigm, routinely practiced in many applications of pattern-recognition algorithms, is centered around evaluating the capacity of already extracted models to derive quantities of interest from new, potentially later encountered individuals. If an already extracted model embodying an identified relationship, reflected in the estimated parameters, is assessed in new individuals whose data were not used to estimate the parameters, the statistical analysis can be said to be an *out-of-sample prediction*. This form of building models from data has been explicitly optimized for and is naturally applicable to a single data point, such as one whole-brain scan or one sequenced genome of a particular individual. Note that we cannot compute the usual p-values on the selected input variables ([23](#_ENREF_23), [24](#_ENREF_24)). This is because the variable selection procedure is itself a random process that is ignored by the theoretical guarantees of classical inference for statistical significance ([22](#_ENREF_22)). Put in yet another way, data-driven model selection is corrupting hypothesis-driven statistical inference because the sampling distribution of the parameter estimates is altered, causing classical statistical to become invalid and the p values become optimistically biased ([22](#_ENREF_22)).

**Simulation**

It is been noted that predictive guarantees are often challenging to derive based on formal theory ([6](#_ENREF_6), [20](#_ENREF_20)). -> empirical simulutations

One place where statistics and computation seem to converge beautifully is when the model is expressed as a simulation: All variables have clear semantic interpretations

**Results**

**Discussion**

The underlying motivation differs, if the canonical linear model is used for inference or prediction.

The statistical paradigms anchored at inference or prediction are common in trying to evaluate whether an effect found in some data extrapolates to another sample of observations drawn from the same underlying population ([7](#_ENREF_7), [15](#_ENREF_15)).

methods common in machine learning can quantify the prediction performance of a previously built algorithm applied to untapped data, such as from a new incoming patient, as a performance metric and immediate practical usefulness.

Tools for statistical hypothesis testing and more recently emerged machine learning techniques can be used to draw different types of conclusions from data. Whereas the core interest of machine-learning applications is to *predict* future events on the basis of patterns observed in data, classical statistics applications are probably more often used to *infer* scientific insight from the effects observed in data ([25](#_ENREF_25)). Both modeling paradigms can serve distinct statistical purposes. Depending on the ultimate clinical or research question, a different set of statistical tools may suggest itself as more appropriate ([26](#_ENREF_26)). It is therefore important for investigators and psychiatrists to acknowledge the partly diverging modeling goals and scopes of interpretation of these two distinct statistical cultures ([2](#_ENREF_2), [27](#_ENREF_27)).

an association between a gene and a psychiatric disorder like schizophrenia with a statistically significant p-value does not necessarily imply that the same gene will be the best choice to successfully predict whether a given individual is affected by schizophrenia. Conversely, an effect that has been empirically shown to be highly predictive of schizophrenia disease based on cross-validation in independent individuals does not always go hand-in-hand with classical statistical tests evaluated to a significant p-value ([27](#_ENREF_27), [28](#_ENREF_28)). For these reasons, *cross-validated machine-learning algorithms and more traditional tools for null-hypothesis testing can sometimes lead to diverging conclusions in certain practical analysis settings (see Fig. 4 for an example).*

**Even a model that fits observed data well can yield poor inferences and predictions about some quantities of interest**

**Breiman2001: what meaning can one give to statements that “variable X is important or not impor- tant.” This has puzzled me on and off for quite a while… variable importance has always been defined operationally. My definition of variable importance is based on prediction. A variable might be considered important if deleting it seriously affects prediction accuracy.  “Importance” does not yet have a satisfactory the- oretical definition**

This statistical goal is in many cases incompatible with the pragmatic wish to somewhat blindly exploit the quantifiable consequences of brain pathophysiology to achieve most accurate *predictions* about the future of individuals based on diverse and rich biological information. Appreciation of this *inference-prediction divergence* will probably be a necessary milestone in personalized medicine research, which will ultimately benefit the well-being of suffering psychiatric patients.

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).

We therefore caution that care needs to be taken when combining both inferential regimes in practical data analysis; We will now illustrate a case of "culture clash" between extrapolation based on classical inference and out-of-sample generalization. The issue has very recently gained momentum

out-of-sample generalization may be particularly important for a future of personalized psychiatry and neurology because cross-validated predictive models can be applied to and obtain answers from a *single patient* (Stephan et al., 2015b). Classical inference by null-hypothesis testing cannot typically produce such *intra-individual predictions* as it is constrained to using the entire data sample to test for (theoretical) extrapolation of an effect at the *population level* (Bzdok et al., 2016b; Arbabshirani et al., 2017).

Neuroscience is entering the era of large-scale data collection, curation, and collaboration (Poldrack and Gorgolewski, 2014) with a pressing need for statistical approaches tailored for the data-rich setting. These may frequently lie beyond the scope of the statistical repertoire cherished today

**Conclusion**

Rivalry between Babylonian and Greek scienctist -> Judea Pearl

Many modelliung tools for inference are rooted in the first half of the 20th century

**A core conviction of classical stats is that: inference is more important than prediction**

**A core conviction of ml is that: prediction is more important than inference!**

for patient-tailored monitoring, risk assessment, and therapeutic intervention.

Awareness of the exposed cultural gap is important to keep pace with the increasing information granularity of acquired neuroimaging repositories.

Ultimately, the modeling goals of inference and predictions are related cousins but they are not twins ([6](#_ENREF_6))

**Acknowledgements**

DB was funded by the Deutsche Forschungsgemeinschaft (DFG, BZ2/2-1, BZ2/3-1, and BZ2/4-1; International Research Training Group IRTG2150), Amazon AWS Research Grant (2016 and 2017), the German National Merit Foundation, as well as the START-Program of the Faculty of Medicine (126/16) and Exploratory Research Space (OPSF449), RWTH Aachen. The authors declare no competing interests.

**Figure Legends**

**Figure X**

****

**Predictability versus significance in four medical datasets.** Integrative plots summarize the inferential importance of each linear-model coefficients (p-values on *x-axis*, log-transformed) and the predictive importance of coefficient sets (out-of-sample R2 scores on *y-axis*, obtained from model application on data not used for model fitting). **A)** The body weight is to be derived from 8 measures in 189 newborns. 3 out of 8 measures are statistically significantly associated with birth weight at p < 0.05 (*red line*). Yet, a predictive linear model explains only 8% of the variance in new babies (R2=0.08). **B)** Prostate specific antigen (PSA), a molecule for prostate carcinoma screening, is to be derived from 8 measuresin 87 men. None of the 8 coefficients reaches statistical significance based on ordinary linear regression, although the fitted coefficients of the predictive model achieve 42% explained variance in unseen men. **C)** Disease progression after one yearto be derived from 10 measures in442diabetes patients. Body mass index (BMI) gives the only significant coefficient (p=0.01), which alone however explains only an estimated 3% of disease progression in future patients.The full coefficients of the predictive model achieve46% explained variance in independent patients. **D)** Lung capacity as indicated by forced expiratory volume (FEV) is to be derived from 4 measuresin 654 healthy individuals. All measures easily exceed the statistical significance threshold. However, a predictive model incorporating body height alone performs virtually on par with predictions based on all 4 coefficients (R2=0.74 versus R2=0.76).

**References**

1. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nature Methods. 2018;15:233–4.

2. Breiman L. Statistical Modeling: The Two Cultures. Statistical Science. 2001;16(3):199-231.

3. White AR. Inference. The Philosophical Quarterly (1950-). 1971;21(85):289-302.

4. Gigerenzer G. The superego, the ego, and the id in statistical reasoning. A handbook for data analysis in the behavioral sciences: Methodological issues. 1993:311-39.

5. Efron B, Tibshirani RJ. Statistical data analysis in the computer age. Science. 1991;253(5018):390-5.

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

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

8. Ioannidis JP. The Proposal to Lower P Value Thresholds to. 005. JAMA : the journal of the American Medical Association. 2018.

9. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute. 2011.

10. Goodfellow IJ, Bengio Y, Courville A. Deep learning. USA: MIT Press; 2016.

11. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504-7.

12. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering. 2018;2(3):158.

13. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:170701836. 2017.

14. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.

15. Casella G, Berger RL. Statistical inference: Duxbury Pacific Grove, CA; 2002.

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

17. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255-60.

18. Bzdok D, Karrer T. Single-Subject Prediction: A Statistical Paradigm for Precision Psychiatry. Brain Network Dysfunction in Neuropsychiatric Illness: Methods, Applications and Implications. New York: Springer; 2018.

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

20. Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms: Cambridge University Press; 2014.

21. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25(6):714-21.

22. Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. The Annals of Statistics. 2013;41(2):802-37.

23. Taylor J, Tibshirani RJ. Statistical learning and selective inference. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(25):7629-34.

24. Loftus JR. Selective inference after cross-validation. arXiv preprint arXiv:151108866. 2015.

25. White AR. Inference. The Philosophical Quarterly. 1971;21:289-302.

26. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: Springer; 2013.

27. Bzdok D. Classical Statistics and Statistical Learning in Imaging Neuroscience. Frontiers in neuroscience. 2017.

28. Shmueli G. To explain or to predict? Statistical science. 2010:289-310.