**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 on the single-patient level. This shift incurs tension between established tools to draw statistical inference on the broader population and emerging machine-learning tools to achieve accurate future predictions for particular individuals. Here, we provide an explicit comparison between classical linear regression that identifies significant contributing factors and learning algorithms that automatically select predictive measures. In artificial data simulations and widespread medical datasets, we quantitatively characterized instances when inference and prediction agree and disagree. While both approaches to defining importance in empirical science often allowed for similar conclusions, we describe divergence in a number of data-analysis settings: variables can turn out to be predictive but not significant, or significant but not predictive. More complete understanding of different ways to reach rigorous conclusions from data will be a prerequisite for generating biomedical knowledge that is reproducible and clinically exploitable.

**Keywords**: scientific discovery | statistical significance | prediction performance | variable importance

# Introduction

Inference and prediction serve different purposes in scientific inquiry in human health and disease. Let’s take diabetes mellitus as an example. The inference paradigm strives to generate insight into the biological causes that lead to high blood sugar levels (hyperglycemia). Diabetes can be a result of insufficient production on insulin hormon in the pancreas (type 1, onset mostly in children). Diabetes may also result from insulin receptors in body cells not responding correctly to the molecule (type 2, onset mostly in adults). Diabetes can further affect previously healthy pregnant women (gestational diabetes). The clinical manifestation of disturbed blood sugar probably underlies partly diverging pathophysiological mechanisms. Distinct biological mechanisms allow for other therapeutic targets. While Type 1 can be treated by injecting missing insulin, type 2 can by counteracted by surgery in some obese patients. Gestational diabetes in turn usually resolves automatically after delivery.

Instead, the prediction paradigm aims to extract regularities to answer what will happen next, here, in a given patient. Diagnosing diabetes in somebody can be achieved based on frequent urination or increased thirst, possibly in combination with age and gender, or some of the many consequences, including retina damage or kidney alterations. Such automatic disease detection is possible without mechanistic understanding of what led to or maintains the disease. Further, a pattern-learning algorithm may recognize missing production of insulin (type 1), or presence of pregnancy in women. However, successful detection of diabetes type alone may shed limited light on the underlying biological basis. Finally, an insulin pump could be engineered that achieves nuanced forecasting of sugar metabolism specific to a particular patient. Again, understanding why aberrant blood glucose levels can be lowered with specific insulin release in a specific individuals may not be necessary to be effective in practice.

, it will be increasingly possible to first quantitatively derive disease stratifications directly from brain measurements in a data-guided fashion to then capitalize on the discovered brain-based phenotypes for patient-tailored monitoring, risk assessment, and therapeutic intervention.

the past 20 years, new technologies (microarrays in genetics + brain imaging in medicine + bag-of-words in finance/marketing) have changed the way that data are collected in fields as diverse as finance, marketing and medicine.

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

Today, single-subject prediction becomes always more feasible due to the recent co-occurrence in data availability, computing power, and cheaper data storage (Goodfellow et al., 2016; Manyika et al., 2011). Brain-scanning and genetic measurements in psychiatry produce massive amounts of data at high granularity that classical statistical tools have not initially been invented to tackle (Efron, 2012). In contrast, machine learning was designed to extract patterns from such observational data that was frequently acquired outside of a carefully controlled experimental context. Additionally, many machine-learning approaches specifically motivated for achieving prediction at scale, such as in thousands of individual subjects or for hundreds of outcomes, as well as when outcome variables are hard or expensive to collect. In precision psychiatry for instance, the accurate prediction of a psychiatric disease, the disease course, or efficacy of treatment options in individual patients is the relevant research goal.

With genetic profiling and an even more extensive variety of phenotyping 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 full access to their electronic health records.

While UKBB is above all a medical dataset and was designed for large-scale population epidemiology

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.

However, it is important to appreciate that the potential immediate gains of the pragmatic goal to identify patterns useful to predict clinical endpoints in complex data does not preclude the longer-term urge for *understanding* the biological nature underlying psychiatric diseases like schizophrenia. Carefully designed, meticulously conducted, and logistically expansive experiments to confirm or reject a-priori verbalized research hypotheses in animals and humans will probably remain a cornerstone to generate neuroscientific insight into mental illness.

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 [(](https://paperpile.com/c/P5MWqG/CxBA+uWuy)variables p >> participants n; [Efron 2012; Efron & Hastie 2016)](https://paperpile.com/c/P5MWqG/CxBA+uWuy). 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.

Some investigators have raised doubt that computing p-values to drawn statistical inference will play the same important role for biomedical research in the 21st century.

Efron {Efron, 2012 #6910}: “It is impressive or worrisome that the multiple testing theory of the 80s still plays a major role in the microarray-era statistical inference.”

Gelman: "Really, any p-value can be viewed as a crude measure of sample size..."

http://andrewgelman.com/2009/06/18/the\_sample\_size/

Ioannidis {Ioannidis, 2018 #7023}: "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."

This potentially unprecedented wealth of longitudinal information can be leveraged to identify early 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 trans-diagnostic stratification in mental health.

the creation, curation, and collaboration of extensive

In the quantitative sciences, the invention and application of statistical tools has always been dictated by changing contexts and domain questions (Efron and Hastie, 2016).

Massive data collection is a game changer in neuroscience (Kandel et al., 2013; Poldrack and Gorgolewski, 2014), and in many other public and private areas (House of Commons, 2016; Jordan et al., 2013; Manyika et al., 2011). There is a growing interest in and pressure for data sharing, open access, and building "big data" repositories (Frackowiak and Markram, 2015; Lichtman et al., 2014; Randlett et al., 2015).

- **What most of us learned as statistics as undergrads at university is from a time when data were rare, expensive/precious** **and experiments were explicitly designed in advance** -> Danilo: not for observational data

primary reason why we cannot rely on data models alone is the rapid change in the nature of statistical problems. The realm of applications of statistics has expanded more in the last twenty-five years than in any comparable period in the history of statistics.

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.

**Methods**

**What do we mean by ‘inference’?**

As the term has been borrowed by various scientific fields to mean different things ([1](#_ENREF_1)), we want to make clear that we adopt the sense common in classical statistics ([2](#_ENREF_2)). 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 generative mechanism 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 modelling 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 man by ‘prediction’?**

Prioritizing insight on 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 ([3](#_ENREF_3), [4](#_ENREF_4)). The investigator wants to automatically extract knowledge of regularities in the world searching through 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 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 limited notion of foretelling future, yet-to-be measured observations ([5](#_ENREF_5)). Prediction has been an important focus of activity in the more recent “machine-learning” community {Breiman, 2001 #4148} and corresponds to how data analysis is often practiced in data-intensive industry.

**Using the linear model for inference**

We want to assess the relative contributions of each of the predictors in explaining Y

A non-signiifcant beta coefficent suggest that the variable can be dropped from the model

Each of them corresponds to the null hypothesis that the beta at hand deviates from zero, whereas the other model coefficients do not

It is aobut confidence intervalls of the betas

**Inference is about the input variables for Breiman**

Model assumed to specify the completey probabilistic structure of how the input measures related to each other, as well as with the output

classical inference is about understanding how the response Y changes as a function of the independent input variables x1, x2, … and it is about these separate input variables that p values are usually computed as evidence for relevance of an effect

mechanisms in the data are assumed to be sufficiently described by means and variances alone as parts of the probability model underlying the dataset at hand ([2](#_ENREF_2))

testing is the ultimate goal

fully specified

In classical statistics, inferential conclusions are drawn by formally testing for the existence of an effect expressed under the null-hypothesis (e.g., a gene is not associated with schizophrenia) in opposition to the alternative hypothesis (e.g., a gene is associated with schizophrenia). The ensuing *p*-value indicates whether data from the subject sample at hand are too extreme to occur under the null hypothesis.

In classical null-hypothesis testing, the p-value is computed on the *entire* data from a particular subject sample in a single process. P-values are commonly obtained from all examined individuals (in-sample) and this quantitative outcome can usually not be used to test for the *same* statistical relationship in a later encountered single individual.

**Using the linear model for prediction**

the confusion thing is that it is the motivation that is utterly different, the maths is the same, there is a key difference in perspective

different procedures for assuring the the conclusions can be trusted

We wish to predict Y from some set of predictor values X

**- a lot of the linear model tools are the same, but the goal is different**

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

Backed up by empirical evaluation

ML is very algorithmic and requires a lot of computation

is typically achieved by identifying relationships in one set of subjects as a function of how these patterns persists in other individuals from a different set of subjects. Here, model parameters are typically estimated on some data while the emerging model is explicitly put to the test in some independent data from unseen individuals (Shalev-Shwartz and Ben-David, 2014)

A probability model is not “required” --> with confidence intervals exceeded or not is not an attractive 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

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

methods common in machine learning can quantify the prediction performance of a previously built algorithm applied to untapped data, such as from a new incoming patient, as a performance metric and immediate practical usefulness. This process of evaluating the prediction performance of learning algorithms is typically performed by a two-step procedure called *cross-validation* (Shalev-Shwartz and Ben-David, 2014). In a first step, the machine-learning algorithm is built on a larger part of the dataset. In a second step, emerging candidate algorithms are evaluated and selected on unused data (Hastie et al., 2001). Because all conditions for independent, identically distributed observations are usually met for the left-out data, the out-of-sample prediction performance on the testing data samples can quantify how likely the same pattern could be detected in future, not yet seen patients.

generalization inference empirically measures the robustness of patterns between typically many variables by testing how well an already fitted model extrapolates to unseen brain measurements (Hastie et al., 2001). 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 (Dwork et al., 2015; Varoquaux et al., 2016). This model assessment is done by a cycle of model fitting on a bigger subset of the available data (i.e., 'training set') and subsequent application of the trained model on the smaller remaining part of data (i.e., 'test set').

the set of fitted model coefficients can be viewed as a hypothesis that is evaluated on empirical data

**if the model cannot make predictions it cannot be falsified, in the sense of the philosopher Karl Popper’s proposal for evaluating hypotheses**,

This analysis paradigm, routinely practiced in many applications of pattern-recognition algorithms, is centered around evaluating the capacity of already extracted models to derive quantities of interest from new, potentially later encountered individuals. If an already extracted model embodying an identified relationship, reflected in the estimated parameters, is assessed in new individuals whose data were not used to estimate the parameters, the statistical analysis can be said to be an *out-of-sample prediction*. This form of building models from data has been explicitly optimized for and is naturally applicable to a single data point, such as one whole-brain scan or one sequenced genome of a particular individual. Whether an obtained model is useful in practice is judged based on its performance in achieving accurate predictions in independent individuals[[1]](#footnote-2). One may view these evaluation practices as more conservative measures when the goal is reliable single-subject predictions in patients admitted to a psychiatry hospital in the future

**Simulation**

It is been noted that predictive guarantees are often challenging to derive based on formal theory ([1](#_ENREF_1), [6](#_ENREF_6)). -> 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 (Casella and Berger, 2002; Efron, 2012).

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

an association between a gene and a psychiatric disorder like schizophrenia with a statistically significant p-value does not necessarily imply that the same gene will be the best choice to successfully predict whether a given individual is affected by schizophrenia. Conversely, an effect that has been empirically shown to be highly predictive of schizophrenia disease based on cross-validation in independent individuals does not always go hand-in-hand with classical statistical tests evaluated to a significant p-value (Bzdok, 2017a; Shmueli, 2010). For these reasons, *cross-validated machine-learning algorithms and more traditional tools for null-hypothesis testing can sometimes lead to diverging conclusions in certain practical analysis settings (see Fig. 4 for an example).*

**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).

**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!**

Ultimately, the statistical goals of inference and predictions are related cousins but they are not twins ([1](#_ENREF_1))

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

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

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

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

5. 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.

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

1. Niels Bohr put this point in the following words: “Prediction is very difficult, especially if it's about the future”. [↑](#footnote-ref-2)