**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

**Examples**

1)blood test: works ,but not specific for the disease

2) advertisement on social media

3) education and student ratings

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.

- **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.

**Methods**

**What we mean by “inference”?**

As the term has been borrowed by various scientific communities to mean different things ([1](#_ENREF_1)), we want to make clear that we adopt the sense of classical statistics {Casella, 2002 #6913}. Inference is aimed at uncovering certain truths about a natural phenomenon of interest by answering whether an effect is likely to exists in the world. The service to science of providing novel insight is here achieved by making explicitly assumptions about the data-generating process. Properties of the underlying generative mechanism are then derived by understanding the way the outcome is affected by various measures of interest. The inference paradigm is aimed at a better understanding the relationship between the input measures in exerting influence on the response variable. In particular, the investigator wants to "identify the few important predictors among a large set of possible variables". This intention explains why historically many statistical approache in the empirical sciences are methods with a linear form, even if the “true” relationship in nature maybe more complicated. The modelling agenda is self-consistent in assuming that the data 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 we man by “prediction”?**

Prioritising insight on intrinsic properties of a natural phenomenon is importantly different from the prediction goal.

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

**emphasis on predictive value in the set of model inputs / smaller concern for what the achieved prediction means for the general population from which the sample was drawn**

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

guessiong unobserved quantities

automatically extract knowledge of regularities in the world

the performance aspect

„accurately modeling the world“ (=prediction);

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

It is about things you have not yet seen. It is however a limited notion of inference. You are just trying to say something about a particular aspect of the world and that is: What will happen next? You do not care so much what is inside the blackbox. You are not trying to say something about the world itself.

Prediction accuracy can capture how well the model’s box can emulate nature’s box as a measure of how well the model can reproduce the natural phenomenon producing the data.

What you really want to do in ML is to model what generalized onto tomorrow’s data.

As yet unjobserbved target vriable, often hard to come by or expansive to comby in practice > model used for predictionn in new individuals whose outcome information we do not yet have

figure out meanginful patterns (or hypotheses) that may have been missed by the human observer

distrubtion-free

ML is very algorithmic and requires a lot of computation

ML more aggressively pursued the computational aspects of data analysis

Prediction paradigm: "**predicting the response for future observations“.**

**- typical setting: X is readily available, but the outputs Y cannot be easily obtained**

- "**f hat is treated as a blackbox**“ -> not necessarily assumptions about the data distributions or the functional form of f

- provided that f hat yields accurate predictions for Y we are not typically concerned with the exact form of f

- the irreducible error / Bayes error rate is a central concern: **predicting Y is also a function of the irreducible error.**There may be variables useful for prediction that we have not measured. Since we did not measure them we cannot use them in the prediction function. For instance, in drug toxicity prediction, the quality of the drug itself may be varying but not captured in X or the patient’s well-being on a given day.

Example questions naturally asked by "modeling for prediction":

- How well can predict the risk for an adverse reaction of a patient to a particular drug?

- Company makes a direct-marketing campaign and wants to know in advance how is likely to respond to a mailing based on demographic variables. Deep understanding of why susceptible individuals can be identified is not necessary.

- Is this house over- or underpriced?

-preferred to talk about the accusracy of models

The relatively recent discipline of Machine Learning,

Backed up by empirical evaluation

**has been an important focus of activity in the “machine-learning” community and corresponds to how statistcs are 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

**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 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

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**,

**Simulation**

It is been noted that predictive guarantees are often challenging to derive based on formal theory ([1](#_ENREF_1), [3](#_ENREF_3)). -> 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.

**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**

**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.

**References**

**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).

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