**An empirical study of 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 biological 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 are putting a premium on accurate predictions for single patients. This shift incurs tension between established tools used to infer statistically significant group effects and emerging machine-learning tools to forecast an individual’s future. Here, we provide a direct comparison of the linear model in identifying significant contributing variables and in searching through the most predictive ones. In artificial data simulations and common medical datasets, we quantitatively characterized instances when statistical inference and pattern recognition concur and diverge. While both modeling approaches allowed for rigorous conclusions, we describe disagreement in 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 | data science | variable importance | artificial intelligence | reproducibility

“Change your statistical philosophy and all of a sudden different things become important” Steven Goodman

# Introduction

Inference and prediction are two sides of a coin / can serve distinct purposes / double-edged sword in the scientific inquiry of human health and disease ([1-3](#_ENREF_1" \o "Bzdok, 2018 #7024)). Let’s take diabetes mellitus as a motivating example. The inference paradigm can be used to establish biological facts that provide insight into the pathways of 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 often 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 encourage 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”/existence/presence 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 impairment. Recognizing symptom combinations is possible without detailed understanding of the biological processes that led to or maintain the disease. Further, a pattern-extraction algorithm may reliably detect diabetes based on lacking production of insulin (type 1) or presence of pregnancy in women. However, the hints on diabetes identified by the algorithm may shed limited light on the biological underpinnings. In treatment, an insulin pump can conceivably be engineered that achieves nuanced forecasting of sugar response regularities specific to the metabolism of a particular patient. Similar individualized profiling may enable risk prognosis and early intervention before onset of symptoms or long-term consequences 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 promote scientific knowledge and we want to know what will happen next.

Inference is intimately linked to statistical null-hypothesis testing and guiding conclusion from data by 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" \o "Gigerenzer, 1993 #5945), [5](#_ENREF_5" \o "Efron, 1991 #4942)), this was a time when data were often rare and expensive to acquire ([4](#_ENREF_4" \o "Gigerenzer, 1993 #5945), [6](#_ENREF_6" \o "Efron, 2016 #6362)). Research experiments were therefore carefully designed in advance and well-controlled. The historical context also explains why classical inference was originally intended for answering research questions in small samples that can be addressed by transparent handpicked statistical models with few knobs to tweak (i.e., parameters) ([7](#_ENREF_7" \o "Efron, 2012 #6910)). Many early statistical inventions were especially attuned to yield understanding of the relationship between a few chosen 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 outcome, classical null-hypothesis testing is still the gold standard today. Some investigators have however cast doubt that computing p-values to draw statistical inference will play an invariably important role for biomedical research throughout the 21st century. John Ioannidis stated ([8](#_ENREF_8" \o "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."

Around the turn of the century, the rapidly increasing availability of whole-genome sequencing and high-resolution body scanning techniques ushered biomedical research into the era of “big data” ([7](#_ENREF_7" \o "Efron, 2012 #6910)). 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 (www.ukbiobank.org). Due to the parallel rise in data availability, computing power, and cheaper data storage ([9](#_ENREF_9" \o "Manyika, 2011 #4150), [10](#_ENREF_10" \o "Goodfellow, 2016 #6717)), the realm of data-analysis has probably expanded more in the last two decades than probably ever before ([7](#_ENREF_7" \o "Efron, 2012 #6910)). Flexible predictive algorithms have been specifically tailored for searching through massive data to extract subtle patterns ([6](#_ENREF_6" \o "Efron, 2016 #6362)). Such predictive pattern-learning approaches promise improved clinical translation of empirically justified single-patient prediction in a fast, cost-effective, and pragmatic manner; which is sometimes viewed as less noble science ([11](#_ENREF_11" \o "Shmueli, 2010 #5944)). Indeed, pioneering studies have leveraged predictive "deep learning" algorithms ([12](#_ENREF_12" \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 ([13](#_ENREF_13" \o "Poplin, 2018 #7026)), ii) detect different heart arrhythmia as well as cardiologists in electrocardiograms from 30,000 patients ([14](#_ENREF_14" \o "Rajpurkar, 2017 #7027)), and iii) diagnose malignant skin cancer as well as dermatologists using almost 130,000 pictures ([15](#_ENREF_15" \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’?**

The term has been borrowed by various scientific fields to indicate different things ([6](#_ENREF_6" \o "Efron, 2016 #6362)). We want to make clear that we adopt the technical meaning common in statistical null-hypothesis testing ([16](#_ENREF_16" \o "Casella, 2002 #6913)). Classical inference is aimed at scientific discovery by trying to uncovering “true” properties of a natural phenomenon of interest. Asking whether an effect is likely to exists in the world is for instance especially suited to ask, “Which gene locations *contribute to* or *are associated* with a disease?” Providing such insight as a service to science is typically achieved by making probabilistic assumptions about how the observed data arose (e.g., bell-shaped Gaussian distribution). The underlying structure of a scientific process is then derived by understanding the way an outcome is affected by a set of input measures. The inference paradigm is especially useful to judge 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 ([16](#_ENREF_16" \o "Casella, 2002 #6913)). The modeling agenda is self-consistent in assuming that the ‘fitted’ model is a sufficient, fully specified summary of the studied phenomena. Often combined with careful experimental control and backed up by formal theory, modeling for inference is how traditional academic statistics have routinely dealt with small to medium data from designed experiments.

**What do we mean by ‘prediction’?**

Ascertaining properties of the inner workings of the phenomenon under study is importantly different from conducting empirical research for the sake of prediction. Here, the emphasis is on accurately modeling the world ([17](#_ENREF_17" \o "Hastie, 2001 #3957), [18](#_ENREF_18" \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 a core metric 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 (e.g., “deep” neural-network algorithms). 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 how the data sample arose from the general population. The ‘trained’ 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 ([19](#_ENREF_19" \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 ([20](#_ENREF_20" \o "Henke, 2016 #6718)).

**Using the linear model for inference**

To assess which variables are statistically significant related to an outcome, we evaluate the strength of evidence based on multiple linear regression. Many statisticians have a preference for evaluating significance by considering several measures in the same model, rather than carrying out simple linear regression based on one independent variable only ([cf. 21](#_ENREF_21" \o "Wu, 2009 #5997)). 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 ([16](#_ENREF_16" \o "Casella, 2002 #6913)). The fitted model is assumed to encapsulate a complete description of how the particular input measures increase or decrease in parallel with each other/jointly 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 ([22](#_ENREF_22" \o "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 impact on explaining the output variable. 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 ([23](#_ENREF_23" \o "Tibshirani, 1996 #5961)). The specified model is almost the same, but the goal is different. Its sparsity constraint is potentially the easiest means to enforce that not all input variables are relevant in a linear model and could be left out ([24](#_ENREF_24" \o "Hastie, 2015 #5915)). We want to identify subsets of the input variables with the strongest effects. 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 mechanism 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 selected model automatically chooses the minimal subset of variables necessary for classifying for instance healthy versus diagnosed individuals. At its extreme, we use beta because as an intermediate step to achieve prediction, not because we care about this parameter itself so much. In other words, we optimize the correctness of prediction, not of the beta coefficients.

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 ([25](#_ENREF_25" \o "Shalev-Shwartz, 2014 #6721)). Explicit model checking was performed by evaluating its expected performance on unknown data using a procedure called *cross-validation* ([25](#_ENREF_25" \o "Shalev-Shwartz, 2014 #6721)). 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 ([17](#_ENREF_17" \o "Hastie, 2001 #3957)). 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. To this end, the cross-validation procedure was 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. As the Lasso does not provide a full least-squares fit due to its shrinkage property, we computed unbiased out-of-sample predictions using ordinary least-squares on the collection of active variables. This allowed us to tear apart the role of shrinking and variable selection in prediction.

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 ([26](#_ENREF_26" \o "Taylor, 2015 #5998), [27](#_ENREF_27" \o "Loftus, 2015 #6152)). 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 ([28](#_ENREF_28" \o "Berk, 2013 #6004)). 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 ([28](#_ENREF_28" \o "Berk, 2013 #6004)).

**Simulations**

It has been noted that predictive guarantees are often challenging to derive based on formal theory ([6](#_ENREF_6" \o "Efron, 2016 #6362), [25](#_ENREF_25" \o "Shalev-Shwartz, 2014 #6721)). Moreover, 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 as the ground-truth is known by design. We therefore conceived “empirical” simulations in which the result cannot be trivially anticipated. Instead of simulating a few hand-selected situations commonly encountered in practice, we rigorously combined distinct scenarios over several dimensions, which yielded 113400 unique simulations.

***The proportion of input variables related to the output****.* We arbitrarily considered 14 proportions from 2.5 percent to 100 percent relevant variables in steps of 7.5.

***The ration of samples to variables***. We implicitly controlled this property by varying the number of samples. We covered the lower range between 50 and 100 samples in steps of 10. This range covers the majority of medical and neuroscientific studies. Between 100 and 2000 we increased the sample size in steps of 100. Finally, we considered the case 10000 and 100000 samples, representing scenarios encountered in recent large-scale datasets such as UK Biobank. However, we, refrained from changing the number of input variables to make the models comparable with regard to the explained variance metric *R^2*.

***Corruption through additive noise.*** We considered the following noise levels, in percent: 0, 50, 100, 200, 500, 1000.

***Multicollinearity between the relevant variables***. We introduced different levels of correlation (rho = 0.5 or 0.9) in either about 50 or 100 percent of the relevant variables. We additionally considered the case of uncorrelated variables matching the model assumptions.

***Pathological transformations.*** Next to undistorted models fitted to normally distributed data, we introduced systematic aberrations from the truth the model can possibly capture by applying nonlinear transformations to about 50 percent of the relevant variables. Among those we considered taking the absolute value, the natural logarithm, the exponential, the square root, the multiplicative inverse as well as polynomials of degree 2-5.

**Results**

*Simulated data*

Abc

For convenience, we refrained from running the analysis pipelines on a local workstation. The simulations were realized using a parallel computing server with 48 Intel Xeon CPUs (1,200 - 2,900 GHz) and 62 GB working memory. 1 week of computation

We made a series of observations...

*Real data*

In addition to the simulated datasets, the same comparison between explanatory modeling and predictive modeling was carried out in a common real-world datasets. The quantitative re-evaluation is presented here for four medical datasets that are frequently used as examples in data-analysis teaching and textbooks (e.g., [17](#_ENREF_17" \o "Hastie, 2001 #3957), [24](#_ENREF_24" \o "Hastie, 2015 #5915)).

In the birthweight dataset, ordinary linear regression was used to evaluate the relation of 8 candidate measures to the body weight of 189 newborn babies. [add multi-collinearity?] The 3 effects that reached statistical significance at p < 0.05 comprised the mother's weight at the last menstrual period (p=0.018, lwt), existing history of hypertension (p=0.012, ht), and presence of uterine irritability (p=0.002, ui). The in-sample model fit amounted to R2=0.141. In the prediction setting, linear models were trained and evaluated on the same data. The best estimate of the explained variance expected in other babies from the same population reached only R2=0.08 (as measured by unbiased out-of-sample prediction accuracy) based on the full set of 8 input measures. After automatically silencing the influence of the age of the mother and number of physician visits during the first trimester (ftv), the remaining 6 active measures still allowed for a prediction performance of R2=0.06. These appeared to be a predictive core subset among the input measures because at 5 out of 8 coefficients the linear model prediction deteriorated to be worse than the average model. Comparing the identification of strongest measures by classical inference and prediction on the birthweight data, a few variables easily reached significance. However, based on the same data, it was challenging to obtain a predictive model with convincing pattern generalization to new data, despite the reasonable sample size.

In the prostate cancer dataset, none of 8 input measures turned out to be statistically significantly associated with prostate-specific antigen (PSA) in 87 men. This molecule is widely used by medical doctors for screening and monitoring of cancer to guide whether or not to surgically remove the prostate gland. Cancer volume (lcavol) was closest to being judged important with p=0.081. In contrast, the estimated prediction accuracy achieved R2=0.42 with 8/8 coefficients, R2=0.42 with 5/8 coefficients, R2=0.38 with 3/8 coefficients, and still R2=0.35 with 2/8 coefficients. Notably, the single most useful measure to predict the PSA concentration in a given man was the cancer volume with an explained population variance of R2=0.25 with 1/8 coefficients (lcavol). That is, despite lacking statistical significance, there were coherent predictive patterns in the data that were reliably extracted across several input variables. The combined input from several variables was required to achieve the highest prediction performances. The prediction approach also detailed that lcavol > svi > lweight carry the most relevant information to forecast a man’s PSA level. The ordered ranking coincided with the absolute beta coefficients obtained using linear regression. In the prostate cancer dataset, in-sample model estimation reverberated with (all three positive) variable importance in out-of-sample prediction performance, but was at odds with the obtained insignificant p-values.

In the diabetes dataset, disease progression after one yearwasto be derived from 10 measures in442patients. In modeling for inference, only the body mass index (bmi) was deemed significant at p=0.01 among all input variables. This single measure, however, only accounted for 3% of explained disease progression in the population in modeling for prediction. Adding the second most predictive variable - s5 - to the linear model with bmi, boosted the prediction accuracy to R2=0.42. Adding more and ultimately all input variables into the model led to small additional improvements in prediction performance (R2=0.46). In fact, s5 showed the highest positive beta coefficient (at the beginning of the regularization path, where small sparsity was imposed) but did not turn out as the final variable remaining in the model. In fact, the coefficient for the s1 measure showed a high absolute weight in the beginning of the path, but is automatically silenced in the middle of it. Summing up the results on the diabetes data, the single significant variable carries negligible information to achieve reliable prediction in new data; only when s5 is incorporated in the predictive model, when suddenly achieve very good predictions in new patients not seen the model.

Finally, in the FEV dataset, the lung capacity captured as forced expiratory volume (FEV) was to be derived from 4 measuresin 654 healthy individuals. All input variables easily successfully exceeded the statistical significance threshold. Yet, a predictive model built on the same data revealed that considering body height alone performed virtually on par with predictions based on all 4 coefficients (R2=0.74 versus R2=0.76). That is, age, gender and smoking habits all easily reached statistical significance, but offered little value for the purpose of prediction. In the case of lung capacity prediction, the predictive variable selection concurred with highest absolute coefficient in both approaches to determined importance. The prediction regime may here miss the potentially mechanistically relevant of influence of smoking by being much more pragmatic. The high significance of all input variables may have been facilitated by the comparably high sample sizes.

**Discussion**

Conducting >100,000 empirical simulations was instructive in providing some quantitative insight into how achieving accurate predictions in new individuals can depart from identifying statistically significant effects across individuals. As our main conclusion, we discovered an asymmetry in how relevant effects are established in modelling prediction and modelling for inference. Throughout a diversity of data analysis scenarios possible in everyday research, statistically significant relationships were not always guaranteed to also enable successful predictions when applying the model to other individuals. Effects robust at the common significance level of p < 0.05 varied between virtually no and almost 100% explained variance in fresh data. By contrast, effects not significant at p < 0.05 mostly failed to deliver useful predictions. In short, even small predictive performances typically coincided with finding underlying significant statistical relationships in almost all cases. However, even statistically strong findings with very low p-values shed only modest light on its value for goal of prediction based on the same data.

Real world settings

Desire to isolate true effects and extending biomedical knowledge

all four possible cases occur in practice:

\* significant and predictive

\* significant but not predictive

\* not significant but predictive

\* not significant and not predictive

IMPORTANCE

Most researchers in biology and medicine face questions of data analysis. What does it mean that a variable is ‘important’ or not? Statistical significance was determined by whether an input measure would take the actually obtained value at least 19 out of 20 times if its impact on the outcome is not important. An official report of the American Statistical Association (ASA) emphasized that “Statistical significance is not equivalent to scientific, human, or economic significance” ([29](#_ENREF_29" \o "Wasserstein, 2016 #6823)). An association between a candidate gene and diabetes grounded in a statistically significant p-value may not necessarily imply that the same gene will be the best choice to successfully predict whether a given individual will be affected by that disease. In a similar vein, there is accumulating evidence from the current replication crisis in psychology that significant results published in a scientific paper are in many cases not substantiated when the identical experiments and data analyses are conducted again at a later point in time ([30-32](#_ENREF_30" \o "Collaboration, 2015 #7032)). The used Lasso method considered variable ‘importance’ in a different way. A variable was considered relevant when leaving it out hurt the ensuing prediction accuracy ([2](#_ENREF_2" \o "Breiman, 2001 #4148)). Some authors believe that such empirical validations to establish importance may increase in the future due to expanding adoption of code and data sharing, as they facilitate across-study and across-method confirmation ([33](#_ENREF_33" \o "Donoho, 2017 #7030)).

In fact, ‘importance’ has probably no uniform theoretical basis ([2](#_ENREF_2" \o "Breiman, 2001 #4148)) and can take different flavors even in the canonical linear model. Just because an approach gives quantitative answers, does not mean that the approach has been the optimal choice for the underlying question by the investigator. Put differently, using p-values or prediction accuracies for backing up claims have both flaws and incomplete in some way ([21](#_ENREF_21" \o "Wu, 2009 #5997), [24](#_ENREF_24" \o "Hastie, 2015 #5915)). This source of uncertainty and misunderstanding begs for intensified research efforts. The ASA statement recommended: "No single index should substitute for scientific reasoning" ([29](#_ENREF_29" \o "Wasserstein, 2016 #6823)) - a viewpoint shared by other prominent investigators ([34](#_ENREF_34" \o "Cohen, 1990 #5949), [35](#_ENREF_35" \o "Gigerenzer, 1987 #6345)). In particular, Ioannidis and colleagues recently monocultural training of biomedical scientists in statistical null-hypothesis testing as one reason behind some of the frequent misuses of p-values ([36](#_ENREF_36" \o "Szucs, 2017 #7029)).

does not always go hand-in-hand with; to back claims; differently nuanced; embrace; irrespective of; informed judgment by the investigator; predictive focus/inference focus; sharpen the distinctino between; explanatory and predictive qualities; set the stage for; predictive modeling/explanatory m.;

**Conclusion**

The present investigation quantitatively exposed how the linear-regression model - a workhorse in many areas of empirical research - can be used for more than one motivation, depending on the ultimate clinical or research question. The more common use of these tools and their extensions to uncover properties of biological processes may give some way to the aim for pragmatic forecasting of clinical endpoints. Care needs to be taken in practical data analysis. Some statisticians have proposed that modeling tools should be defined by the problems they can be applied to solve, rather than cataloguing methods under particular umbrella terms ([37](#_ENREF_37" \o "Friedman, 2001 #5937)). It is important for investigators and clinicians to acknowledge the partly diverging modeling goals and scopes of interpretation of different modelling agendas ([2](#_ENREF_2" \o "Breiman, 2001 #4148), [38](#_ENREF_38" \o "Bzdok, 2017 #6436)). Statistical literacy may become increasingly important for taking rigorous and reproducible steps on our way to personalizing medical care, which will ultimately benefit the well-being of suffering patients.

The prediction-inference distinction may also remind us of some of Claude Bernard’s ideas ([39](#_ENREF_39" \o "Bernard, 1957 #7028)). Prediction may be closer to what he called empirical medicine oriented towards practical patient care as an often theory-free endeavor, such as symptom monitoring, risk assessment, and choosing therapeutic intervention. Statistical inference may bear a more direct relationship to his conceptualization of scientific medicine aimed at elucidating unknown principles underlying biological processes driven by theory, such as asking for the reasons why certain individuals are at risk for disease onset or illuminating why a certain drug works better in some of them.

It may increasingly become apparent that the modeling goals of inference and prediction, even when using a linear model and using the same data, should be viewed as related cousins but not twins ([6](#_ENREF_6" \o "Efron, 2016 #6362)). Awareness of the strength and weakness of both "data-analysis cultures" is important to avoid missing critical information and to keep pace with the accelerating data deluge in biomedicine.

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

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

13. 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.

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

15. 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.

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

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

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

19. 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.

20. 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.

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. Gelman A, Hill J. Data analysis using regression and multilevelhierarchical models: Cambridge University Press New York, NY, USA; 2007.

23. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267-88.

24. Hastie T, Tibshirani R, Wainwright M. Statistical Learning with Sparsity: The Lasso and Generalizations: CRC Press; 2015.

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

26. 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.

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

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

29. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129-33.

30. Collaboration OS. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716.

31. Feynman RP. The Meaning of It All: Thoughts of a Citizen-Scientist. Reading: Addison-Wesley. 1998.

32. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nature methods. 2015;12(3):179.

33. Donoho D. 50 Years of Data Science. Journal of Computational and Graphical Statistics. 2017;26(4):745-66.

34. Cohen J. Things I have learned (so far). American psychologist. 1990;45(12):1304.

35. Gigerenzer G, Murray DJ. Cognition as intuitive statistics. NJ: Erlbaum: Hillsdale; 1987.

36. Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. Frontiers in human neuroscience. 2017;11:390.

37. Friedman JH. The role of statistics in the data revolution? International Statistical Review/Revue Internationale de Statistique. 2001:5-10.

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

39. Bernard C. An introduction to the study of experimental medicine: Courier Corporation; 1957.