**PLOS Call: Machine Learning in Health and Biomedicine**

**Prediction and inference diverge in biomedicine:**

**Simulations and real-world data**

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

In the 20th century many advances in biological knowledge and evidence-based medicine were grounded in p-values and accompanying methods. In the beginning 21st century, ambitions towards precision medicine put a premium on detailed predictions for specific individuals. The shift causes tension between established methods used to infer statistically significant group differences and burgeoning machine-learning tools used to forecast an individual’s future. Here, we offer a direct comparison of the linear model for identifying significant contributing variables and for searching the most predictive ones. In artificial data simulations and common medical datasets, we elaborated how statistical inference and pattern recognition can concur and diverge. Across all cases, even small predictive performances typically coincided with finding underlying significant statistical relationships. 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. More complete understanding of different ways to define ‘importance’ is a prerequisite for reproducible research findings that can be 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 when inquiring human health and disease ([1-3](#_ENREF_1" \o "Bzdok, 2018 #7024)). Let’s take diabetes mellitus as a motivating example. The inference paradigm is effective to establish biological details that provide insight into the pathways leading 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 reflect 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 may encourage other therapeutic interventions that have been shown to yield statistically significant benefits for a particular patient group. Especially patients with type 1 diabetes can profit from injecting missing insulin, while type 2 diabetes can sometimes be alleviated by surgery in obese patients. In turn, diabetes developed in the pregnant patient group usually resolves after delivery without treatment intervention.

Instead of substantiating the presence of effects in disease biology and clinical treatment, the prediction paradigm aims to detect statistical regularities that hold in the future. Diabetes can be diagnosed “superficially” based on frequent urination or increased thirst, possibly combined with age and gender, or its secondary consequences like retina damage or kidney impairment. Recognizing such 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 insufficient production of insulin (type 1) or pregnancy in women. However, the disease hints identified by the algorithm may shed little light on the biological underpinnings. In treatment, it can be possible to engineer an insulin pump that achieves accurate forecasting of nuanced sugar response regularities specific to a patient’s metabolism. Similar individualized monitoring may enable risk prognosis and early intervention before onset of symptoms or longer-term consequences to improve medical care, without requiring understanding the precise biological pathways at play. In this way, inference and prediction have important but different contributions to make to biomedical research - we want to promote scientific knowledge of disease and we want to know what will happen next at the individual level.

Classical inference has been intimately linked to statistical null-hypothesis testing and drawing conclusions from data using p-values. This framework emerged in the first half of the 20th century ([4](#_ENREF_4" \o "Cowles, 1982 #4363), [5](#_ENREF_5" \o "Cox, 2006 #7037)) together with tools like linear regression, *t-*tests, and ANOVA. Electrical calculators not yet widely available ([6](#_ENREF_6" \o "Gigerenzer, 1993 #5945), [7](#_ENREF_7" \o "Efron, 1991 #4942)), this was a time when data were rare and expensive to acquire ([6](#_ENREF_6" \o "Gigerenzer, 1993 #5945), [8](#_ENREF_8" \o "Efron, 2016 #6362)). Hence, research experiments were often carefully designed in advance and well-controlled. The historical context also explains why classical inference was originally intended for answering research questions in subjects recruited to the local laboratory that can be addressed by transparent statistical models with few knobs to tweak (i.e., model parameters) ([9](#_ENREF_9" \o "Efron, 2012 #6910)). Many early statistical inventions were especially tuned to yield understanding of the relationship between a few handpicked candidate measures. Many of today’s medical doctors and biomedical researchers have been “raised” with this statistical culture at the university. If the scientific 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 ([10](#_ENREF_10" \o "Wasserstein, 2016 #6823)). However, a few investigators, including John Ioannidis, have cast doubt that computing only p-values to draw statistical inference will continue to play an invariably important role for biomedical research ([11](#_ENREF_11" \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 imaging ushered biomedical research into the era of “big data” ([9](#_ENREF_9" \o "Efron, 2012 #6910), [12](#_ENREF_12" \o "Blei, 2017 #7035), [13](#_ENREF_13" \o "Leonelli, 2016 #6996)). 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 data (e.g., nutrition, lifestyle, medications) for extensive phenotyping of 500,000 volunteers - the currently largest biomedical data resource of its kind (www.ukbiobank.org). Due to the parallel rise in data availability, computing power, and cheap data storage ([14](#_ENREF_14" \o "Manyika, 2011 #4150), [15](#_ENREF_15" \o "Goodfellow, 2016 #6717)), the realm of data-analysis has expanded more rapidly in the last two decades than perhaps ever before ([9](#_ENREF_9" \o "Efron, 2012 #6910)). Flexible predictive algorithms are particularly suited for searching through massive data to extract subtle patterns ([8](#_ENREF_8" \o "Efron, 2016 #6362)). Such predictive modeling approaches promise improved clinical translation of single-patient prediction in a fast, cost-effective, and pragmatic manner. The goal of empirically justified predictive success is sometimes viewed as a less noble science ([16](#_ENREF_16" \o "Shmueli, 2010 #5944)). Nevertheless, pioneering "deep learning" studies ([17](#_ENREF_17" \o "Hinton, 2006 #5956)) have recently demonstrated the potential to 1) predict the cardiovascular risk, blood pressure, and smoking behavior from retina scans using medical data from almost 300,000 patients ([18](#_ENREF_18" \o "Poplin, 2018 #7026)), 2) detect different heart arrhythmia as well as cardiologists in electrocardiograms from 30,000 patients ([19](#_ENREF_19" \o "Rajpurkar, 2017 #7027)), and 3) diagnose malignant skin cancer as well as dermatologists using almost 130,000 pictures ([20](#_ENREF_20" \o "Esteva, 2017 #6829)).

It is important to appreciate that the potential immediate gains of the pragmatic goal to exploit relationships predictive for clinical endpoints in complex medical data does not preclude the scientific curiosity to better understand the primary biology of diseases like diabetes. Carefully planned and expensive experiments to confirm or reject a-priori verbalized research hypotheses in animals and humans will probably remain a cornerstone to generate biomedical knowledge. In this computational investigation, we try to bring common approaches to classical inference and pattern prediction to the same table to illuminate their characteristic commonalities and differences.

**Methods**

**What do we mean by ‘inference’?**

The term has been used by several quantitative fields with varying definitions ([8](#_ENREF_8" \o "Efron, 2016 #6362)). Here we adopt the technical meaning common in statistical null-hypothesis testing ([21](#_ENREF_21" \o "Casella, 2002 #6913)). Such classical inference is aimed at scientific discovery by trying to uncover “true” properties of the studied phenomenon. Quantifying whether an effect exists in the world is especially suited to ask scientific questions like ‘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., the bell-shaped Gaussian distribution). The underlying structure of a scientific process is typically derived by understanding the way a set of input measures affect an outcome. The inference paradigm is especially useful to judge the individual relevance of each quantitative measure in impacting the response of interest. In particular, the investigator commonly wants to quantitatively isolate the more important predictors among the set of candidate variables, which were often hand-selected based on previous research. This intention explains why historically many empirical sciences have long relied on linear model approaches, even if the “true” relationship in nature may be more complicated ([21](#_ENREF_21" \o "Casella, 2002 #6913)). Modeling for inference is self-consistent in assuming that the ‘fitted’ model is a sufficient, fully specified summary of the studied phenomena, with each variable have a clear semantic interpretation. Often combined with careful experimental control and backed up by formal theory, this modeling agenda is how traditional academic statistics has routinely dealt with small to medium data from planned data acquisition.

**What do we mean by ‘prediction’?**

Seeking properties of the inner workings of the phenomenon under study can depart from empirical research for the sake of prediction. To accurately model the world, the investigator here wants to automatically extract knowledge of regularities searching through possibly meaningful patterns ([22](#_ENREF_22" \o "Hastie, 2001 #3957), [23](#_ENREF_23" \o "Jordan, 2015 #5958)). This modeling goal is for instance especially suited to ask, ‘Which gene locations are *useful* to *detect* whether an individual has a disease or not?’ Prediction accuracy is an established metric to capture how well the quantitative model can *emulate* a high-level description of mechanisms in nature; that is, how well the built model can reproduce the studied phenomenon that has been quantitatively measured in the data. In the extreme case, the quantitative model may embody the discovered statistical relationship in a way that is opaque to the investigator (e.g., many “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 ([24](#_ENREF_24" \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 industries ([25](#_ENREF_25" \o "Henke, 2016 #6718)).

**Using the linear model for inference**

To assess which variables have a statistically significant relation to the outcome, we evaluated the strength of evidence using ordinary linear regression. Many statisticians have a preference for assessing significance by considering several measures in the same model, rather than carrying out simple linear regression based on one independent variable at a time (cf. [26](#_ENREF_26" \o "Wu, 2009 #5997), [27](#_ENREF_27" \o "Freedman, 1983 #6539)). A single input variable can turn out to be insignificant by itself, but become significant when part of a model with other input variables ([28](#_ENREF_28" \o "Hastie, 2015 #5915)). This probably most common approach to perform ordinary least-squares regression optimizes the following objective:

where is the number of individuals who contributed to the dataset, is the number of input variables (called *independent, explanatory* or *predictor 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. Given that the other variables are also present in the model, 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 ([21](#_ENREF_21" \o "Casella, 2002 #6913)). The fitted model is assumed to encapsulate a description of how the particular input measures increased or decreased in parallel with each other to jointly explain variability in the response of interest.

After model estimation, statistical inference was drawn to decide whether the contribution of input variable in explaining the response is sufficiently important to be deemed *statistically* *significant*. The relevance of the effects is computed based on the confidence intervals of the beta coefficients ([29](#_ENREF_29" \o "Gelman, 2007 #7004)). Inferential conclusions are drawn by formally testing for deviance of the observed effects from the null-hypothesis (e.g., a gene is not associated with schizophrenia) in line with the alternative hypothesis (e.g., a gene is associated with schizophrenia). The ensuing p-values for the input variables indicated whether our data provided enough evidence against the null hypothesis of no relevant relationship. The approach attempted to reject the null hypothesis that the beta coefficients are truely zero, with no relation to the response variable. A non-significant beta coefficient suggests 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 classical linear regression, we chose a variant of linear regression as a minor extension to use it as a predictive pattern-learning algorithm ([30](#_ENREF_30" \o "Tibshirani, 1996 #5961)). LASSO also estimates a linear model, but the goal revolves around prediction. Its sparsity constraint is arguably the simplest method to enforce that not all input variables are relevant in the linear model and each has the same chance to be left out in the final model to predict based on new observations ([28](#_ENREF_28" \o "Hastie, 2015 #5915)). We thus wanted to identify subsets of the input variables with the strongest predictive effects. Automatic variable selection was achieved by minimizing the same optimization objective augmented with a penalty term:

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 to be predicted (called *target variable*) by expressing it 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 during model fitting. The higher the higher the tendency to set specific coefficients to exactly zero, which effectively “silences” the corresponding measure’s influence on explaining the output variable. An explicit probability model is not required - whether the confidence intervals exceeded a threshold or not is here often no optimality criterion for variable importance. This approach did also not assume that means and variances fully describe the probabilistic mechanism in the data, only that they are informative enough to make useful predictions about the future. Once fitted, the model was applied to other samples to *predict* unobserved outputs or ”shipped” to other laboratories for repeated application. The selected model thus automatically chose the minimal subset of predictive variables necessary for classifying for instance healthy versus diagnosed individuals. At its extreme, many pattern-learning models use the coefficient estimates as an intermediate step to achieve prediction, less because we cared about the parameter values themselves. In other words, many predictive modeling approaches prioritize the correctness of the prediction on new data, rather than the estimation of particular beta coefficients.

Following model estimation, the importance of the candidate predictive model was evaluated based on the cross-validation gold standard to obtain explicit empirical guarantees {Hastie, 2001 #3957}. Answering the question whether an obtained predictive algorithm generalizes to unseen data is tackled in an empirical 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 were estimated on some data while the emerging model is explicitly put to the test in some independent data from unseen individuals ([31](#_ENREF_31" \o "Shalev-Shwartz, 2014 #6721)). Explicit model checking was performed by evaluating its expected performance on unknown data ([31](#_ENREF_31" \o "Shalev-Shwartz, 2014 #6721)): First, the linear model was built on a larger part of the dataset. Second, emerging candidate algorithms were evaluated and selected on unused data to avoid an overly optimistic evaluation of goodness-of-fit ([22](#_ENREF_22" \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. In this way, the cross-validation scheme quantified the so-called *out-of-sample performance* as 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 common modification helped us to isolate the influence of shrinking and variable selection in forming predictions with LASSO. As an important consequence, all prediction scores reported in this work were obtained from ordinary least-squares regression based on the full set or subset of input variables automatically selected from the preceding LASSO estimation.

Such modeling for prediction, 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. 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 automatically selected input variables ([32](#_ENREF_32" \o "Taylor, 2015 #5998), [33](#_ENREF_33" \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 ([34](#_ENREF_34" \o "Berk, 2013 #6004)). Put differently, data-driven model selection is corrupting hypothesis-driven statistical inference because the sampling distribution of the parameter estimates is altered, causing classical statistical inference to become invalid and the p-values would have become optimistically biased ([34](#_ENREF_34" \o "Berk, 2013 #6004)).

**Simulations**

It has been noted that formal guarantees for the expected model prediction performance are often challenging to derive by mathematical theory ([8](#_ENREF_8" \o "Efron, 2016 #6362), [31](#_ENREF_31" \o "Shalev-Shwartz, 2014 #6721)). In such settings, empirical simulation can come to the rescue for studying the properties of statistical methods in controlled computational experiments ([29](#_ENREF_29" \o "Gelman, 2007 #7004)). Here we explicitly confronted linear modeling for inference and for prediction in a series of synthesized datasets, columns of input variables each related or not related to the outcome . We generated each dataset based on a set-up ground-truth model , where are the fixed random coefficients, X is a matrix with samples and variables with random entries drawn from a standard Gaussian distribution and denotes the added Gaussian noise. Each dataset was fed into linear models (cf. above) with the aim to identify significant input measures or to identify input measures most useful for accurate predictions on new observations. To sharpen the distinction between explanatory and predictive approaches in general, we systematically varied distinct aspects of the data-generating process.

1. Model violations**:**To examine more closely how inference and prediction behave when the linear model is known not to capture how the data came about, we introduced pathological transformations on 50% of the relevant variables in *X*. In addition to datasets with exclusively linear effects, deviations between the generating and fitting model were introduced by taking the absolute value, the natural logarithm, the exponential, the square root, the multiplicative inverse, as well as polynomials of degree 2-5.
2. Proportion of informative variables:To study how the fraction of informative versus unrelated variables modulate the inferential and predictive processes, we varied the proportion of non-zero coefficients in the model for generating . We considered 14 proportions ranging from only one to all input variables carrying information about the response .
3. Samples-to-variables ratio: To investigate the relation between the number of samples relative to the number of variables *p*, we systematically varied the number of available observations. We covered the lower range between 50 and 100 samples in steps of 10, which probably well reflects a majority of medical and neuroscientific studies. Between 100 and 2,000 samples we increased the sample size in steps of 100. Moreover, we considered the extreme cases 10,000 and 100,000 samples, which cover recent large-scale datasets such as the UK Biobank. The total number of input variables was kept constant to preclude secondary effects on the results due to changing model capacity.
4. Signal-to-noise ratio:To explore the role of nuisance variation in the data, such as induced by imperfect measurement techniques, we systematically increased the noise in how the ground-truth model relates to the response . The noise terms was therefore multiplied by 0 (i.e., canceling out any noise), 0.5, 1, 2, 5, or 10.
5. Redundant versus unique sources of information: To elucidate how correlated input measures trade-off against each other with respect to the outcome, we introduced different degrees of pairwise covariation between the variable columns of (i.e., multicollinearity). Ground-truth models also generated data from a multivariate Gaussian distribution that exposed 50% or 90% percent of common variation between the relevant variables, besides datasets containing only independent variables (i.e., 0% covariation).

The collection of simulated datasets amounted to 113,400 different data-analysis settings. For each of them, we focused on the best (smallest) p-value from the coefficient and the highest out-of-sample prediction performance as quantified by the score. All simulations were carried out on a parallel computing server with 48 Intel Xeon CPUs (1,200 - 2,900 GHz) and 62 GB of working memory. The simulations required almost 4 weeks of computation time and produced 2 GB of modeling results.

**Scientific computing implementation**

Python was selected as the scientific computing engine. Capitalizing on its open-source ecosystem helps enhance replicability, reusability, and provenance tracking. The *statsmodels* package was used to estimate ordinary least squares regression and obtain p-values (http://statsmodels.github.io). The *scikit-learn* package ([35](#_ENREF_35" \o "Pedregosa, 2011 #4211)) provided efficient, unit-tested implementations for handling state-of-the-art machine learning algorithms ([http://scikit-learn.org](http://scikit-learn.org/)). All analysis scripts that reproduce the results of the present study are readily accessible and open for reuse (<http://github.com/banilo/to_be_added_later)>.

**Results**

*Simulated datasets*

Across 113,400 randomly synthesized datasets, we made a series of observations about the characteristic differences that emerge in a direct comparison of seeking statistical inference and assessing model prediction are. Fitting linear models to series of datasets generated with increasing non-linear effects easily reached significance but distinctly varied in the predictability of the outcomes (Fig. 2B; Fig. 3). It was expected that even, as opposed to odd, polynomial transformation (e.g., x2 or x4) incur larger violations to model validity because the direction of effects in the input variables is lost. As such, 4th-order polynomial incurred worse model fit than 5th-order expansion, entailing both worse p-values and R2 out-of-sample prediction performance.To emulate random variation such as from measurement error, we added gradually increased noise in the data. This additional challenge during model fitting decreased the predictability more systematically than the significance (Fig. 2C). More adverse effects in the data to be analyzed were not observed to entail less models with variables that turned out statistically significant.To emulate the frequently encountered challenges due to multi-collinearity, we have increased the correlation shared between the input measures (Fig. 2D). More variation common to several input variables appeared to worsen the p-values more than the prediction performance. Covariance of 90% yielded (smallest) p-values closer to the typical p = 0.05 threshold and seldom extremely low p-values. Concurrently, many data-analysis scenarios that did not yield a single significant relation between an input variable and the response of interest were generated in this high-noise setting. To capture some implications of the ongoing trend to data aggregation and data accumulation in biomedicine, we gradually increased the number of available data points per generated dataset (Fig. 2E). At the highest sample size of n=10,000, low significance tended to more systematically agree with low predictability and extremely high significance also concurred with almost perfect out-of-sample performance. That is, in datasets, bigger than is currently the norm, we eventually observed more consistent co-occurrence of significance and prediction.Exploring the proportion of relevant measurements in the ground-truth model (Fig. 2F), we noted that fewer truly relevant inputs gave rise to strongly significant p-values in the presence of poor predictive performance. Finally, applying models that did not correspond to the data-generating process of the input and output variables (Fig. 2A) led to results with high significance and predictions in many cases. However, using the valid (linear) model to fit the randomly generated data allowed for many of the best prediction performances (Fig. 3A).

*Real medical datasets*

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., [22](#_ENREF_22" \o "Hastie, 2001 #3957), [28](#_ENREF_28" \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?] In the widely used multiple regression approach, 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 LASSO automatically silenced 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 strongest measures identified by classical inference and prediction on the birthweight data, a few input 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, we suddenly achieved very good predictions in new patients not seen by 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 the 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**

Analyzing more than 100,000 empirical simulations and several biomedical datasets offered some 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 modeling prediction and modeling for inference. Charting 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.

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 most 1 in 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’ ([10](#_ENREF_10" \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, in psychology and other empirical sciences ([36-39](#_ENREF_36" \o "Collaboration, 2015 #7032)), there is accumulating evidence from a replication crisis 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. The used predictive 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 practices, as they facilitate across-study and across-method confirmation ([40](#_ENREF_40" \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. A statistical method that produces automatic importance judgments still requires informed judgment how far the conclusions can be trusted - the initial choice of method may be more or less optimal for the underlying research question. Put differently, using p-values or prediction accuracies for backing up claims have both flaws and each is incomplete in some way ([26](#_ENREF_26" \o "Wu, 2009 #5997), [28](#_ENREF_28" \o "Hastie, 2015 #5915)). The ASA statement recommended: ‘No single index should substitute for scientific reasoning’ ([10](#_ENREF_10" \o "Wasserstein, 2016 #6823)) - a viewpoint shared by other prominent investigators ([41](#_ENREF_41" \o "Cohen, 1990 #5949), [42](#_ENREF_42" \o "Gigerenzer, 1987 #6345)). In particular, Ioannidis and colleagues recently stressed monocultural training of biomedical scientists in statistical null-hypothesis testing as one reason behind some of the frequent misuses of statistical methods ([43](#_ENREF_43" \o "Szucs, 2017 #7029)).

**Conclusion**

The present quantitative investigations exposed how models for linear regression - a workhorse in many areas of empirical biomedical research - can be used for more than one motivation. The more common use of these tools and their extensions to uncover properties of biological processes may be supplemented for pragmatic forecasting of clinical endpoints. Some statisticians therefore proposed that analysis tools should be defined by the modeling goals they can be applied to achieve, rather than cataloguing methods under umbrella terms, such as ‘exploratory’ vs. ‘confirmatory’ or ‘statistics’ vs. ‘data science’ ([40](#_ENREF_40" \o "Donoho, 2017 #7030), [44](#_ENREF_44" \o "Friedman, 2001 #5937)). It is important for investigators and practicing medical doctors to acknowledge the partly diverging modeling agendas and ensuing scopes of interpretation ([2](#_ENREF_2" \o "Breiman, 2001 #4148), [45](#_ENREF_45" \o "Bzdok, 2017 #6436)). Statistical literacy may become increasingly relevant 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 ([46](#_ENREF_46" \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.

In transitioning towards a future of precision medicine, it is increasingly critical that mainstream statistics and machine learning are related but different, even when the data are the same and widespread linear models are used ([8](#_ENREF_8" \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**

We thank several investigators for insightful comments on a previous version of the manuscript: B.T. Thomas Yeo (National University of Singapore), Nikolaus Kriegeskorte (Columbia University), Benjamin de Haas (University of Giessen), and João Sato (Universidade Federal do ABC).

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. DE acknowledges support by the Amazon AWS Research Grant (2015), the German National Merit Foundation, as well as the French National Institute for Informatics and Automation (INRIA) (Starting Researcher Position SRP 2016). The authors declare no competing interests.

**Figure Legends**

**Figure 1**

****

**Predictability versus significance in simulated datasets.** Based on 113,400 different simulations, the discrepancy between explanatory and predictive modeling was quantified in a wide range of possible data-analysis settings. The generated variables and outcome were fed into linear models with the goal to draw classical inference (smallest p-value among all model coefficients, x axis) and to evaluate model forecasting performance on never seen data (out-of-sample R2 score of the model, y axis). **A)**Hexagonal binning summarizes how many simulations led to a particular prediction-inference relation area-by-area in a 2D histogram. This visualization technique was proposed for aggregating data with a high number of observations ([47](#_ENREF_47" \o "Carr, 1987 #7038)). **B)**Statistical significance and prediction accuracy are juxtaposed, exposing relation to the commonly applied p < 0.05, p < 0.01, and p < 0.001 thresholds (bigger grey circle means bigger sample size). In the large majority of conducted data analyses, at least one input variable was significantly related to the response variable at p < 0.05 (red dashed vertical line). However, based on the same data, we observed considerable dispersion in how well significant models were able to make useful predictions on fresh data points.

**Figure 2**

****

**Properties underlying analysis results of simulated data.** Disentangles how and when linear modeling for significance testing (smallest p-value, x axis) and linear modeling for prediction (out-of-sample R2 score, y axis) agreed and diverged across our constructed datasets. **A)** Pathological settings, where the chosen model does not correspond to the data-generating process of the input and output variables, tended to yield better significance and predictions. **B)** Fitting a linear model to data with increasing non-linear effects easily reached significance but distinctly varied in predictability of outcomes.**C)** Increasing random variation in the data, which can be viewed as emulating measurement errors, appeared to decrease the predictability more systematically than the significance.**D)** Increasing correlation between the input measures appeared to worsen the p-values more than the prediction performance.**E)** Increasing the number of available data points eventually yielded occurrences of strong significance and prediction.**F)** Small numbers of relevant predictors allowed for scenarios with highly significant p-values in combination with poor predictive performance.

**Figure 3**

****

**Implications of different model pathologies in simulated data.** Consequences of applying a linear model to data that contain non-linear mechanisms of different types and degrees (cf. Fig. 2A). Certain non-linear effects are likely to influence measurements of various real biological systems. That is, some misalignment between the data and the commonly employed linear model is likely to be the rule rather than the exception.

**Figure 4**

****

**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, using the linear model for prediction explained 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 reached statistical significance based on ordinary linear regression, although the fitted coefficients of the predictive model achieved 42% explained variance in unseen men. **C)** Disease progression after one yearisto be derived from 10 measures in442diabetes patients. Body mass index (BMI) gave the only significant coefficient (p=0.01), which alone however explained 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 quantified by forced expiratory volume (FEV) is to be derived from 4 measuresin 654 healthy individuals. All measures easily exceeded the statistical significance threshold. However, a predictive model incorporating body height alone performed virtually on par with predictions based on all 4 coefficients (R2=0.74 versus R2=0.76). In sum, linear models can show all combinations of predictive vs. not and significant vs. not in every-day data analysis.

**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. Cowles M, Davis C. On the Origins of the .05 Level of Statistical Significance. American Psychologist. 1982;37(5):553-8.

5. Cox DR. Principles of statistical inference: Cambridge university press; 2006.

6. 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.

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

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

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

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

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

12. Blei DM, Smyth P. Science and data science. Proceedings of the National Academy of Sciences. 2017;114(33):8689-92.

13. Leonelli S. Data-centric biology: a philosophical study: University of Chicago Press; 2016.

14. 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.

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

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

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

18. 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.

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

20. 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.

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

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

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

24. 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.

25. 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.

26. 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.

27. Freedman DA. A note on screening regression equations. the american statistician. 1983;37(2):152-5.

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

29. Gelman A, Hill J. Data analysis using regression and multi-level hierarchical models: Cambridge University Press New York, NY, USA; 2007.

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

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

32. 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.

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

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

35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research. 2011;12:2825-30.

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

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

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

39. Ioannidis JP, Khoury MJ. Improving validation practices in “omics” research. Science. 2011;334(6060):1230-2.

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

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

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

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

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

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

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

47. Carr DB, Littlefield RJ, Nicholson W, Littlefield J. Scatterplot matrix techniques for large N. Journal of the American Statistical Association. 1987;82(398):424-36.