Patterns

Inference and Prediction Diverge in Biomedicine

Highlights

- We systematically juxtapose variable selection using significance versus prediction
- Successful prediction often coincided with significant p values
- Yet strong statistical significance did not always coincide with predictive value
- Understanding the inference-prediction dilemma is imperative for precision medicine

Authors

Danilo Bzdok, Denis Engemann, Bertrand Thirion

Correspondence

danilo.bzdok@mcgill.ca

In Brief

In systematic empirical data simulations and real-world biomedical datasets, we explore scenarios of agreement and disagreement between variables identified as statistically significant or variables identified as predictively relevant.

Bzdok et al., 2020, Patterns *1*, 100119
November 13, 2020
https://doi.org/10.1016/j.patter.2020.100119

Patterns

CellPress
OPEN ACCESS

Article

Inference and Prediction Diverge in Biomedicine

Danilo Bzdok,[1,2,5,6,*] Denis Engemann,[3,4] and Bertrand Thirion[3]
[1]Mila – Quebec Artificial Intelligence Institute, Montreal, QC, Canada
[2]Department of Biomedical Engineering, McConnell Brain Imaging Centre (BIC), Montreal Neurological Institute (MNI), Faculty of Medicine, School of Computer Science, McGill University, Montreal, QC, Canada
[3]INRIA Saclay, CEA, Université Paris-Saclay, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France
[4]Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[5]Twitter: @danilobzdok
[6]Lead Contact
*Correspondence: danilo.bzdok@mcgill.ca
https://doi.org/10.1016/j.patter.2020.100119

THE BIGGER PICTURE Across research communities, the analysis goals of inference and prediction are two sides of a coin. Many empirical studies leaning on statistical significance typically focus interpretation on the best p values obtained for one or a few variables. In contrast, many empirical studies dedicated to prediction are backed up by cross-validated model performance on fresh data points.

In a future of single-patient prediction from big biomedical data, it may become central that modeling for inference and modeling for prediction are related but importantly different. The relevant subset of variables identified based on p values or based on predictive value can converge or diverge depending on the data scenario. We show that diverging conclusions can emerge even when the data are identical and when widespread linear models are used. Awareness of the relative strengths and weaknesses of both "data-analysis cultures" may become unavoidable in navigating between complementary goals in scientific inquiry.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

In the 20th century, many advances in biological knowledge and evidence-based medicine were supported by p values and accompanying methods. In the early 21st century, ambitions toward precision medicine place a premium on detailed predictions for single individuals. The shift causes tension between traditional regression methods used to infer statistically significant group differences and burgeoning predictive analysis tools suited to forecast an individual's future. Our comparison applies linear models for identifying significant contributing variables and for finding the most predictive variable sets. In systematic data simulations and common medical datasets, we explored how variables identified as significantly relevant and variables identified as predictively relevant can agree or diverge. Across analysis scenarios, even small predictive performances typically coincided with finding underlying significant statistical relationships, but not vice versa. More complete understanding of different ways to define "important" associations is a prerequisite for reproducible research and advances toward personalizing medical care.

> "Change your statistical philosophy and all of a sudden different things become important"
>
> Steven Goodman

INTRODUCTION

Across research communities, the analysis goals of inference and prediction are two sides of a coin in the scientific inquiry of human health and disease.[1,2] The overarching goal of the present study was to directly compare the result of common linear modeling practices for prediction or for inference in the most typical way encountered in everyday biomedical data analysis. Many empirical studies centered on statements backed up by p values therefore typically focus interpretation on the lowest (i.e., best) p values obtained for one or few variables; whereas many empirical studies revolving around prediction backed up by cross-validation cherish aggregate metrics of explained variance that index performance of the

Box 1. What Do We Mean by "Inference"?

The term has been used by several quantitative fields with varying, sometimes conflicting definitions.[10] Here we adopt the technical meaning common in the statistical null-hypothesis testing context.[29] Statistical inference, typically involving regression models, point estimation, and hypothesis testing, is aimed at scientific discovery by trying to reveal "true" properties of the studied phenomenon. Quantifying whether an effect exists in the world is especially suited to ask scientific questions such as "Is a genetic polymorphism *associated with* or *has an effect on* 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 explored by trying to understand 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 affecting the response of interest. The investigator wishes to draw inference by quantitatively isolating the more important measures among the set of candidate variables, which were often hand-chosen based on existing knowledge. This intention explains why, historically, many empirical sciences have long relied on linear model approaches, even if the "true" relationship in nature is thought to be more complicated.[29] Modeling for inference is self-consistent in assuming that the "fitted" specified model is a sufficient summary of the studied phenomena, whereby each variable and its units have an immediate semantic interpretation. Often combined with careful experimental control and formally backed up by mathematical theory, the inference agenda is how traditional academic statistics has routinely dealt with small to medium datasets from planned data acquisition.[10]

model as a whole. Let us take diabetes mellitus as a motivating example.

The inference paradigm (Box 1), often realized by classical two-sample hypothesis testing, is effective in establishing biological effects that provide some insight into what leads to disturbed blood sugar levels at the population level. Diabetes in children can be a result of insufficient production of insulin hormone in the pancreas (type 1). Diabetes in adults may also reflect deficient insulin receptor response in body cells (type 2). Diabetes can moreover affect previously healthy pregnant women (gestational type). The clinical manifestation of disturbed blood sugar probably underlies partly diverging pathophysiology. Such differences in biological pathways may encourage other therapeutic interventions that have been shown to yield statistically significant benefits for a particular patient group. Long-practiced inferential statistics can also substantiate clinical observations that most patients with type 1 diabetes profit from injecting missing insulin, while obese patients with type 2 diabetes are more likely to profit from surgical intervention; and symptoms in pregnant patients usually resolve after delivery.

Instead of trying to demonstrate the presence of group effects in disease biology and clinical treatment, the prediction paradigm (Box 2) aims to detect statistical regularities that generalize to the other data,[3–6] including other patients and the future of the same patients. Diabetes can be detected and diagnosed "superficially" by pattern-recognition algorithms based on frequent urination or increased thirst, possibly combined with age and sex/gender, or its later downstream consequences such as retina damage or kidney impairment. Recognizing such symptom constellations is possible without requiring detailed mechanistic understanding of the biological processes that led to or maintain the disease. In treatment, analysis tools with a focus on prediction can make it possible to engineer an insulin pump for accurate forecasting of the sugar response regularities that characterize a patient's metabolism. Similar individualized predictive monitoring may enable forecasting somebody's risk to be affected by a given disease, and may enable early intervention before onset of symptoms or longer-term consequences to improve medical care, without requiring understanding the metabolic mechanisms at play. In this way, the general analysis goals of inference and prediction make important but partly distinct contributions to biomedical research: We want to extend scientific knowledge of disease in general, and we want to know what may happen next to a particular individual.

The inference agenda has been intimately linked to statistical null-hypothesis testing and drawing conclusions from data guided by p values. This framework of ongoing importance emerged in the first half of the 20th century[7] for use with tools such as linear regression, t tests, and ANOVA. This was a time when electrical calculators were not yet widely available,[8,9] when data were rare and expensive to acquire.[8,10] Hence, research experiments were and still are often well controlled and carefully designed in advance. 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).[10,11] Many early statistical inventions were intended to yield understanding of the relationship between a few candidate measures that were handpicked, guided by the scientific question and previous research. Many of today's medical doctors and biomedical investigators have been "raised" with this statistical culture. If the scientific goal is to examine whether an effect exists or which specific input variables have most impact on an outcome, null-hypothesis testing legitimately remains the gold standard.[12] However, a few authors, including John Ioannidis, have cast doubt on the notion that computing p values and the accompanying methodology to draw statistical inference will continue to play an invariably important role for biomedical research.[13] One recurring reaction to mitigate publication bias emanating from the widespread practice of filtering based on significance testing has been a ban on p values by certain journal editorial boards or strong encouragement of reporting accompanying effect sizes and standard deviations (e.g., *The Journal of Basic and Applied Social Psychology*).[14]

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."[11,15,16] There is growing momentum for the creation and curation of massive datasets. For instance, the UK Biobank has gathered

Box 2. What Do We Mean by "Prediction"?

Describing aspects of the inner workings of the studied phenomenon (cf. Box 1) is conceptually distinct from empirical research for the purpose of pattern recognition. To accurately model the world in this way, the investigator wants to extract knowledge of regularities searching through possibly meaningful candidate patterns.[25,47] This modeling goal is for instance especially suited to ask "Is there a set of genetic polymorphisms *useful* to *detect* whether a disease is present or not at the level of single individuals?" Compared with modeling for inference (at the group level), there tends to be relatively less focus on small details of the data-collection process, also because useful predictions are often "found" rather than obtained from carefully planned experimental studies. Prediction accuracy is the core metric to capture how well the quantitative model can encapsulate 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 strives for highly accurate guesses by explicitly checking the fitted model by external validation. The "trained" quantitative model is built for prediction in new individuals whose outcome information we would only obtain in the future. Typically, the predicted outcomes cannot be easily obtained, are expensive, or are otherwise hard to come by.[38] This aspect of automatically "filling in" missing information also explains why mere correlation between two variables, such as in Pearson's correlation coefficient, may represent a more limited notion of foretelling yet-to-be measured observations.[3] While out-of-sample prediction can be performed with a variety of regression analysis tools, this modeling goal has been an important focus of activity in machine-learning and other communities,[1] and corresponds to how data analysis is often practiced to solve problems in data-intensive industries.

genetic, behavioral, environmental, and lifestyle data for extensive phenotyping of 500,000 volunteers— currently the largest biomedical data resource of its kind (www.ukbiobank.org). Due to the parallel improvements in data availability, computing power, and data storage,[17,18] the realm of data analysis has probably expanded faster in the last two decades than ever before.[11,15] Flexible prediction algorithms are particularly well suited for sieving through rich data to extract subtle patterns.[10] Such modeling approaches aimed at prediction can be less transparent but promise improved clinical translation of single-patient prediction in a fast, cost-effective, and pragmatic manner. It may be closer to engineering than science when investigators place special emphasis on the success of empirical predictions rather than formal modeling properties or how to unlock biomedical insight.[19] Nevertheless, pioneering studies have now demonstrated the potential of "deep-learning" algorithms in medicine.[20] Deep-learning tools were shown to work well to: (1) predict the cardiovascular risk, blood pressure, and smoking behavior from retina scans using medical data from almost 300,000 patients;[21] (2) detect different heart arrhythmia as well as cardiologists in electrocardiograms from 30,000 patients;[22] and (3) diagnose malignant skin cancer as accurately as dermatologists using almost 130,000 pictures.[23]

There is tremendous potential in the practical goal to exploit predictive relationships to forecast clinical endpoints from medical data. However, such empirical success may not fully satisfy the scientific curiosity to understand the primary biology of diseases such as diabetes. Carefully planned and expensive experiments to confirm or reject a priori verbalized research hypotheses in animals and humans will surely remain a cornerstone for generating biomedical knowledge.

In our systematic empirical investigation, we therefore try to bring analytical tools aimed at classical inference and pattern prediction to the same table, which are usually separately employed, to illuminate their characteristic commonalities and differences. Our data-simulation approach allows precise control over and, thus, careful comparisons between important dataset dimensions ranging from (1) available sample size, to (2) number of informative input variables and (3) redundancy of information carried in the input variables about the outcome, as well as to (4) random noise variation and (5) mis-specification of the quantitative model. These key aspects have strong influence on whether investigators can identify the truly relevant variables from data, and whether modeling for inference versus prediction are prone to tell different stories about the same dataset. Our approach is summarized in Figure 1.

RESULTS

Simulated Datasets: Synoptic Summary

Across 113,400 constructed datasets (Figure 2), we observed several characteristic differences between seeking statistical inference and maximizing model prediction. 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 (Figures 3F and 4). It was expected that even, as opposed to odd, polynomial data transformation (e.g., x^2 or x^4) incur larger violations to model validity because the direction of effects in the input variables is lost. As such, fourth-order polynomial expansion deteriorated model fit more than fifth-order expansion, entailing both worse p values and worse R^2 prediction performance (out-of-sample). To emulate random variation such as from measurement error, we added gradually increased noise to the data. This additional challenge during model fitting decreased predictability more systematically than significance (Figure 3D). Adding more random noise to the data was not observed to entail fewer models with statistically significant variables. To emulate the frequently encountered challenges when facing collinear data, we have increased the correlation shared between the input measures (Figure 3C). More variation that is correlated across several input variables appeared to worsen p values more than prediction performance. Covariance of 90% yielded p values (i.e., smallest in the model) closer to the typical $p < 0.05$ threshold and seldom very low p values. Concurrently, many data-analysis scenarios that did not yield a single significant relation between an input variable

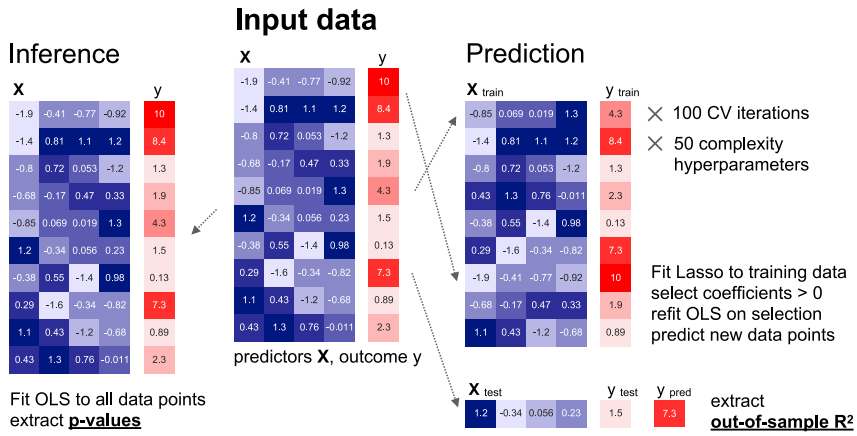

Figure 1. Problem Statement and Schematic of Workflow

Biomedical research has long relied on regression-type analysis tools. However, the same regression analysis tools are often used with non-identical objectives within different scientific communities. In carefully designed experiments in animals and humans, studies have focused on making progress toward inferring the role of preselected variables in explaining the observed outcome or experimental conditions, such as isolating cancer-related genetic variants in mouse models. Propelled by larger datasets and recent advances in machine-learning algorithms, applied clinical research has shifted toward combining heterogeneous and rich measurements from different sources to "brute-force" forecasting of practically useful endpoints, such as predicting the duration of hospitalization of new patients based on previous electronic health records. For a long time, these disparate uses of identical analytical tools have been carried out in parallel with little crosstalk between communities. As a result, the conceptual and empirical relationship between the established agenda of statistical inference and the now expanding agenda of raw prediction performance remains largely obscure. Motivated by this increasing need, our study carried out a careful comparison of modeling for inference (left) and prediction (right) on identical datasets, based on comprehensive empirical simulations and revisiting common medical studies. OLS = ordinary least squares; Lasso = least absolute shrinkage and selection operator.

and the response of interest were generated in this high-collinearity setting. To capture key implications of the ongoing trend to data aggregation in biomedicine, we gradually increased the available data points per generated dataset (Figure 3A). At the highest sample size of n = 100,000, low significance tended to more systematically agree with low predictability, and extremely high significance also mostly concurred with perfect out-of-sample performance. That is, in datasets bigger than is currently the norm, we observed more consistent correspondence between significance and prediction. Exploring different proportions of relevant measurements in the ground-truth model (Figure 3B), we noted that fewer truly relevant inputs gave rise to strongly significant p values in the presence of poor predictive performance. Finally, applying linear models that deviate from the data-generating process of the input and output variables (Figure 3E) led to results with high significance and predictability in many cases. However, using the valid (linear) model to fit the randomly generated (linear) data allowed for many of the best prediction performances (Figure 4A). Importantly, after correcting p values for multiple comparisons using Bonferroni's method, we made the same observations and the ensuing patterns of results remained identical.

Across the 113,400 synthetic datasets, our results show a variety of incongruencies in identifying important variables among a set of candidate variables. Using linear models, statistical significance and accurate predictions showed diverging patterns of success at detecting those variables that we knew were relevant for the outcome. For each dataset simulation, we have therefore computed the recovery based on the subset of ordinary least-squares (OLS) coefficients correctly detected to be significant and the active (i.e., non-zero) coefficients of the Lasso model indeed predictively relevant for the outcome (Figure 5). This metric allowed us to compare the number of correctly identified variables, analogously for OLS and Lasso. Even considering the simulation scenarios with consistent model specification (Figure 5A), systematic disagreements between inference and prediction emerged in a number of simulation cases, depending on available sample size and the number of relevant variables. Datasets in which OLS performed worse than Lasso had in common that the number of data points was larger and the number of relevant variables smaller (horizontal row in Figure 5A). Note that input variables irrelevant for the outcome constitute noise, even if no explicit noise is added to the predictors or the outcome. This can explain why OLS can commit errors if noise is not intentionally added (Figure 5A). Instead, datasets in which OLS performed better than Lasso were characterized by a small to moderate number of data points available for model building (vertical column in Figure 5A). In data-simulation scenarios including misspecified models (Figure 5B), the patterns of disagreements in recovery performance were more diverse. Lasso performed better with smaller sample sizes and larger number of truly relevant variables (upper left in Figure 5B). OLS tended to be more successful at recovering important variables with larger sample sizes and smaller numbers of relevant variables (lower right in Figure 5B).

As such, carefully comparing Lasso and OLS in a variety of synthetic datasets, we observed the largest disagreements in variable identification with small to moderate sample sizes, which is still a very common situation in day-to-day data analysis in biomedicine. The collection of findings shows that more consistent agreement between using linear models for prediction versus inference was observed when the sample size was >1,000 data points (fuchsia to red on diagonal in Figure 5B). Yet even with $\geq$10,000 available data points we report certain disagreements (orange to yellow in Figure 5B).

Simulated Datasets: True Negatives, False Positives, False Negatives, and True Positives

Well-Specified Models

Encouraged by the revealed divergences, we explored deeper explanations by dissecting the variable picks into proportions of true negatives, false positives, false negatives, and true positives (Figures 5, 6, and 7). Overall, in scenarios where the model specification was coherent with the simulated dataset, we

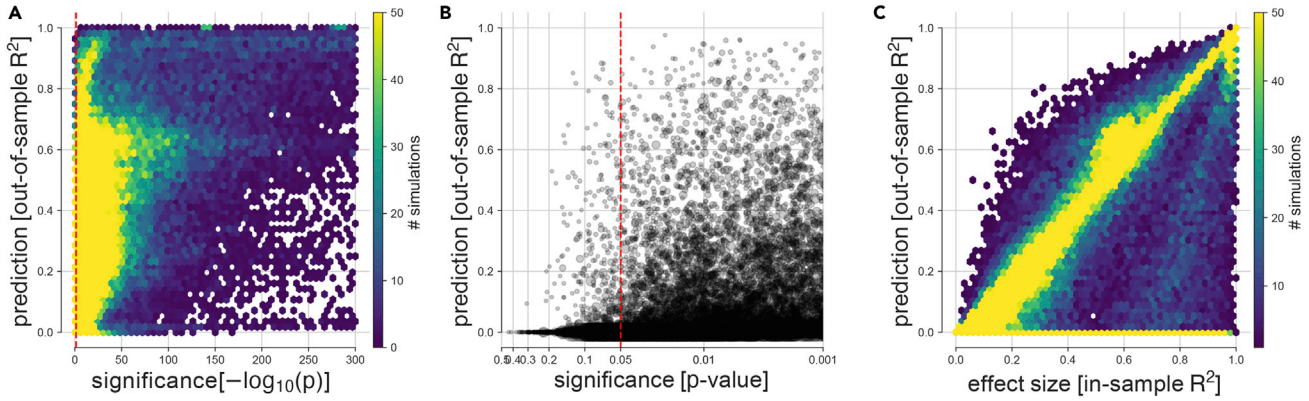

Figure 2. Predictability versus Significance of Effects in Simulated Datasets

(A) Based on 113,400 simulations, the discrepancy between predictive and explanatory modeling was quantified in a wide range of possible data-analysis cases. The generated variables and outcomes were analyzed by linear models with the goal to draw classical inference (single best p value among all model coefficients, x axis) and to evaluate model forecasting performance on never-seen data (out-of-sample R^2 score of the whole model, y axis). We then devised multiple visualizations to grant an overview on complementary facets of this collection of simulation results. Hexagonal binning summarizes how many simulations led to a particular relation between achieved results from prediction and inference in a 2D histogram across a wide range of p values. This area-by-area visualization was proposed for aggregating data with many observations.[24]

(B) Predictive accuracy and statistical significance are juxtaposed at a refined scale with their relation to the commonly applied thresholds at $p < 0.05$, $p < 0.01$, and $p < 0.001$ (bigger gray circle means bigger sample size).

(C) As smaller p values do not necessarily represent stronger statistical evidence, prediction accuracy is compared with the effect size derived from the explained variance on the model-fitting data (in-sample R^2 score of the model). 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 such significant linear models were able to make useful predictions on fresh observations. Note that the smallest p values fall within the range below the smallest 64-bit floating point number around 2.2×10^{-308} that can be represented in Python. Such p values are conceptually plausible and can occur in large simulated datasets such as when many input variables explain the outcome, and noise is absent. For a detailed breakdown of the relationship between inference and prediction according to experimental factors of the simulation, see Figures 3 and 4.

observed disagreements especially for true-negative hits and false-positive errors (Figures 6A and 6B). In false-positive cases, large disagreements between prediction and inference went hand in hand with moderate negative Spearman correlation between the variables that OLS and Lasso picked as important as well as 30% more hits for Lasso than OLS on average (Figure 6B). Lasso showed a specific tendency to commit more false-positive errors in the weak-sample regime (Figure 6B, vertical column), where OLS made fewer erroneous picks in variable detection. On the other hand, Lasso committed fewer false-positive errors than OLS in the strong-sample regime (Figure 6B, horizontal row).

For true-negative hits, we observed the second-largest disagreement in scenarios with benign model specifications given 30% more hits for Lasso on average and only moderate Spearman correlation (Figure 6A). Coherently, the proportion of true negatives depended on the number of relevant variables. That is, with more truly relevant variables in a given dataset fewer true-negative hits can occur. Lasso tended to perform better with a decreasing number of relevant variables. In such cases, OLS tended to perform more poorly. Finally, for false-negative errors and true-positive hits, we observed agreement between Lasso and OLS in our simulations across weak- and strong-sample regimes.

Mis-specified Models

In scenarios where the employed quantitative models could be at odds with the process that has generated the dataset, distinct patterns of disagreement emerged between prediction and inference (Figure 7). Here, the strongest disagreements between Lasso and OLS surfaced for false-positive errors, characterized by negligible Spearman correlation and 8% more false positives for Lasso than for OLS on average (Figure 7B). In a large subset of simulations, Lasso incurred up to about 90% of false positives, where OLS rarely committed more than 10% of these errors. Such observations also occurred in datasets with larger sample sizes. In the other three scenarios (Figures 7A, 7C, and 7D) we noticed consistent disagreements, which were relatively less pronounced. OLS generally detected more true negatives, scored fewer true-positive hits, and committed more false-negative errors. For true-negative hits (Figure 7A), OLS successfully scored fewer hits than Lasso across different numbers of relevant variables. This disagreement increased with smaller sample sizes (left-to-right horizontal gradient above the diagonal of Figure 7A, upward-pointing triangles). At the same time, Lasso also performed worse than OLS in various simulated datasets (right-pointing triangles in lower right). For false-negative errors (Figure 7C), OLS committed more false-negative errors than Lasso in various simulations (rightward-pointing triangles in lower right). This tendency was more pronounced when the number of relevant variables was higher, and escalated in the weak-sample regime. Finally, for true-positive hits (Figure 7D), Lasso scored better than OLS by 10% on average. Here, the advantage of Lasso over OLS was most apparent in the low- and medium-sample-size regimes.

Key Experimental Factors Driving Results

In each of the four outcomes from variable selection, we then sought to understand the global factors that are most responsible for the models' success and failure. We carried out

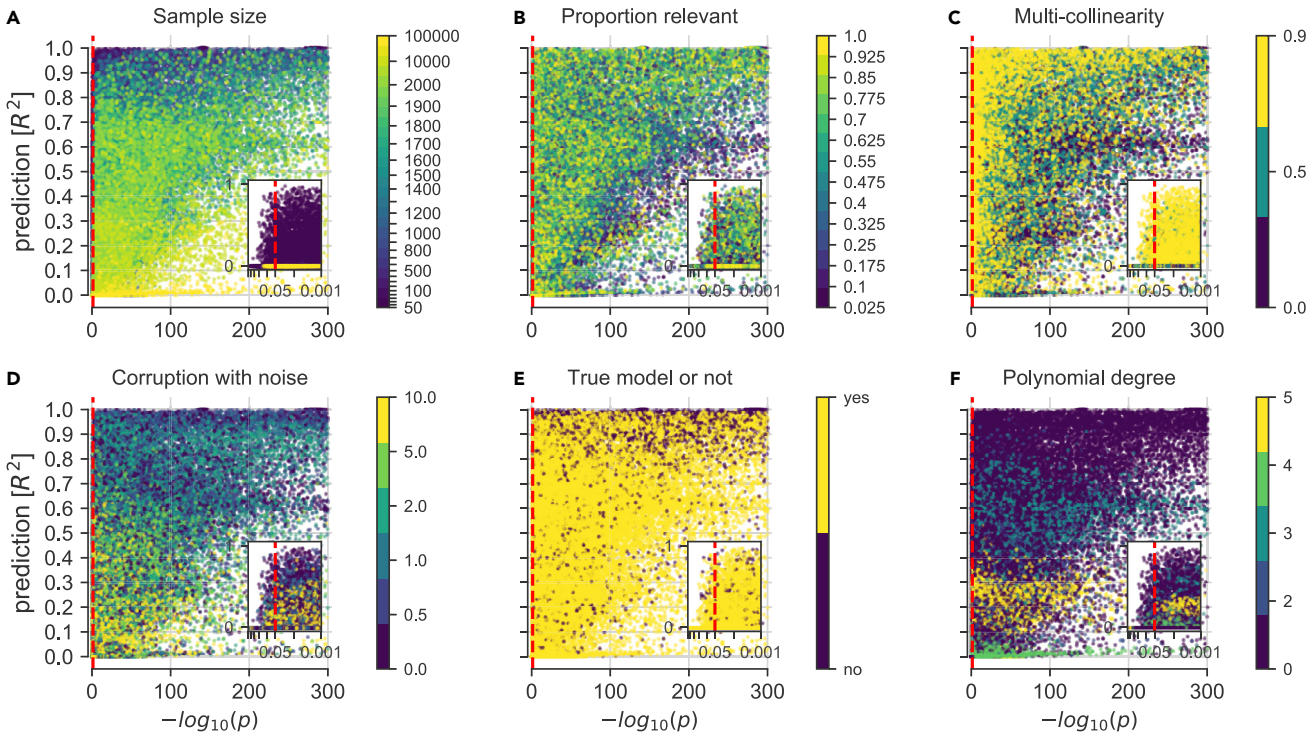

Figure 3. Properties Underlying Analysis Results from Simulated Data

A more detailed exploration of how linear modeling for significance testing (single best p value among all model coefficients, x axis) and linear modeling for prediction (out-of-sample R^2 score of the whole model, y axis) agreed and diverged across constructed datasets.

(A) Increasing the number of available data points eventually yielded co-occurrences of strong significance and prediction.

(B) Small numbers of relevant predictors allowed for scenarios with highly significant p values in combination with poor predictive performance.

(C) Increasing correlation between the input measures, common in biological data, appeared to worsen the p values more than the prediction performance.

(D) Increasing random variation in the data, which can be viewed as imitating measurement errors, appeared to decrease the predictability more systematically than the significance.

(E) Pathological settings, where the chosen model does not correspond to the data-generating process of the collection of input and output variables, can enhance both significance and predictions.

(F) Fitting a linear model to data with increasing non-linear effects easily reached significance but distinctly varied in predictability of outcomes.

Random Forest (RF) regression to summarize performance variation in OLS and Lasso as a function of the experimental properties of the generated datasets (Figure 8). With the possibility to accommodate more complicated relationships, the variable importance metric of the RF allowed us to identify which experimental factors had the largest impact on differences in recovery performance. First and foremost, the number of truly relevant variables in the datasets had a strong impact on explaining variation in model behavior (Figure 8C). Especially for false-negative errors, the number of relevant variables was considerably more explanatory for OLS than Lasso. Second, model violations turned out to be important to account for differences between analyses using OLS or Lasso (Figures 8B–8D). As our third most influential experimental factor, the number of data points stood out as an aspect of the simulated datasets that differentially explained model performance comparing OLS and Lasso.

This principled aggregation from 113,400 synthetic datasets provides a synoptic perspective that strengthens our core observation: Frequently, variable identification in OLS and Lasso was affected in distinct ways by the properties of the data-generating mechanisms.

Real Medical Datasets

To complement the simulated datasets, we carried out the same direct comparison between explanatory modeling and predictive modeling in common real-world datasets (Figure 9). The quantitative re-evaluation is presented here for four medical datasets that are frequently used as examples in data-analysis teaching and textbooks.[25,26]

Birthweight Dataset

Standard linear regression was used to evaluate the relation of eight candidate measures to the body weight of 189 newborn babies (Figure 9A). In this approach, the three 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 $R^2 = 0.141$. In the prediction setting, linear models were trained and evaluated involving the same data. The best estimate of the explained variance expected in babies that we would see in the future reached only $R^2 = 0.08$ (as measured by unbiased out-of-sample prediction) based on the full set of eight input measures. After predictive variable selection "silenced" the influence of the age of the mother and the number of physician

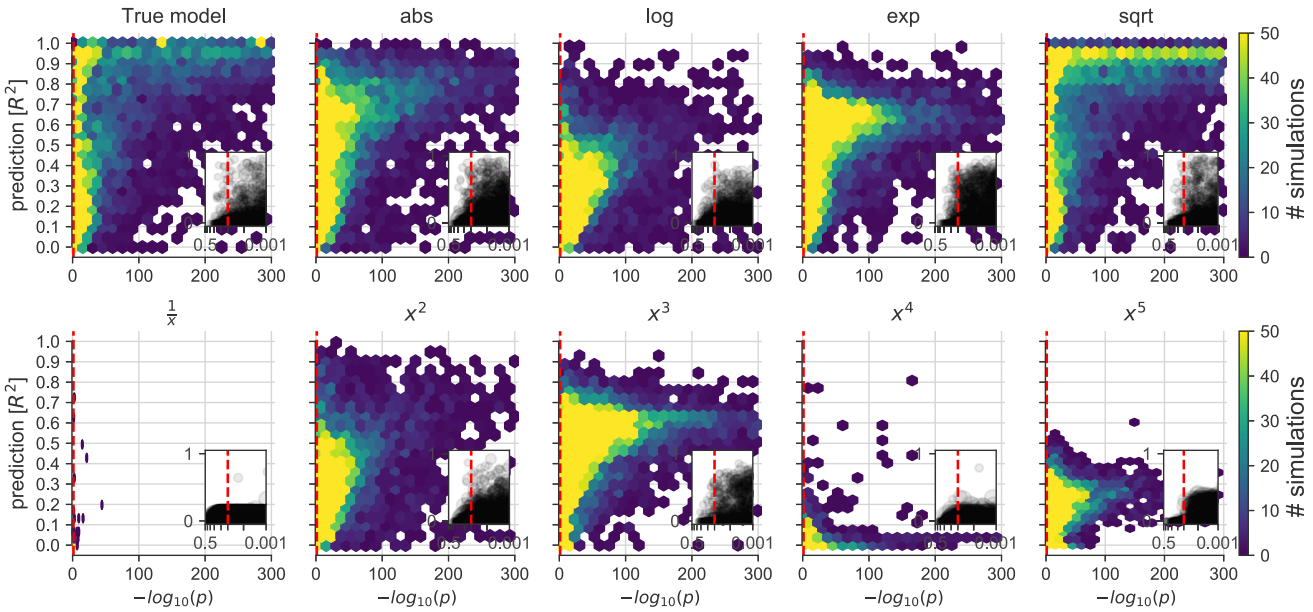

Figure 4. Implications of Different Model Violations in Simulated Datasets

An exploration of the consequences of applying a linear model to datasets that are known to contain non-linear data mechanisms of different types and degrees (cf. Figure 2F). Certain non-linear effects are likely to influence measurements of various real biological systems. That is, in everyday data analysis, some misalignment between the data and the commonly employed linear model is likely to be the rule rather than the exception.

visits during the first trimester (ftv), the remaining six active measures still allowed for a prediction performance of R^2 = 0.06. These appeared to be a predictive core subset among the input measures because at five out of eight coefficients the linear model prediction diminished to be worse than the average model. Comparing the strongest measures identified by classical inference and pattern prediction by explicit model checking on the birthweight data, a few input variables easily reached significance. However, relying on the same data, it was challenging to obtain a predictive model with convincing pattern generalization to new data points, despite the reasonable sample size.

Prostate Cancer Dataset

None of eight input measures turned out to be statistically significantly associated with prostate-specific antigen (PSA) in 87 men (Figure 9B). This molecule is widely used by medical doctors for cancer screening and monitoring 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 R^2 = 0.42 with 8/8 coefficients, R^2 = 0.42 with 5/8 coefficients, R^2 = 0.38 with 3/8 coefficients, and still R^2 = 0.35 with 2/8 coefficients. Notably, the single most useful measure to predict a man's PSA concentration in these data was the cancer volume with an explained population variance of R^2 = 0.25 with 1/8 coefficients (lcavol). That is, despite lacking statistical significance, we found coherent predictive patterns in the data that were reliably extracted. The combined information from several variables was required to achieve the higher 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 β 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 in disagreement with the obtained insignificant p values.

Diabetes Dataset

Disease progression after 1 year was to be derived from 10 measures in 442 patients (Figure 9C). 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 when modeling for prediction. Adding another predictive variable, s5, to the linear model with bmi enhanced the prediction accuracy to R^2 = 0.42. Adding more and ultimately all input variables into the model led to small additional improvements in prediction performance (R^2 = 0.46). In fact, s5 showed the highest positive β coefficient (at the beginning of the regularization path, where small sparsity constraint was imposed) but did not turn out as the final variable remaining in the model. 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 were very good predictions achieved in new patients not yet witnessed by the model.

FEV Dataset

Finally, the lung capacity captured as forced expiratory volume (FEV) was to be derived from four measures in 654 healthy individuals (Figure 9D). All input variables easily reached the statistical significance threshold, but a predictive model built from the same data revealed that considering body height alone performed virtually on par with predictions based on all four coefficients (R^2 = 0.74 versus R^2 = 0.76). That is, age, sex/gender, and smoking habits all easily reached statistical significance but offered little added value for the purpose of prediction. In the

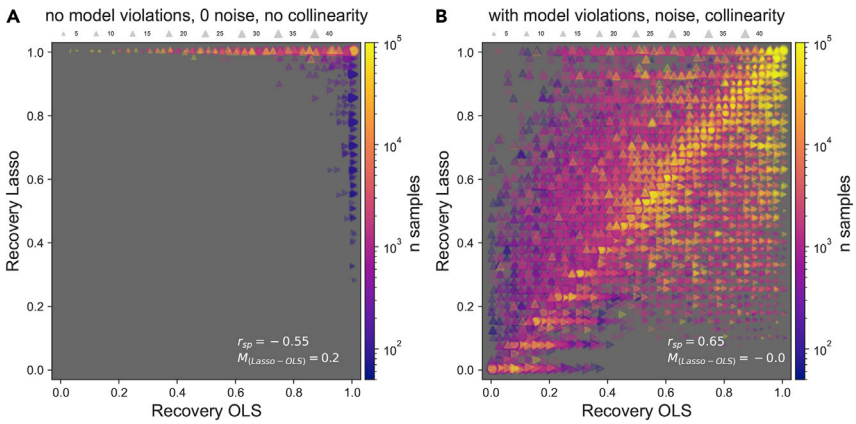

Figure 5. Inference and Prediction Systematically Disagree on Recovering Relevant Variables in Specific Data Scenarios

Across dataset simulations, we assessed the accuracy at which the predictively relevant coefficients of the Lasso or the significantly relevant coefficients ($p < 0.05$) of the OLS captured the known true set of data-generating variables. Circles indicate equal performance and triangles diverging performance. Upward-pointing triangles indicate that the Lasso scored better at variable identification. Right-pointing triangles indicate that OLS performed better. The size of the triangles and markers indicates the number of relevant variables to be recovered. Hotter color indicates higher number of available data points. To measure overall agreement or disagreement, we computed Spearman's rank correlation (r_{sp}) between the accuracies of Lasso and OLS as well as the mean of their differences $M_{(Lasso-OLS)}$. Jittering mitigated overplotting for display of simulation density. In scenarios shown in (A), the linear model was well specified. In scenarios shown in (B), model violations occurred given the presence of non-linear effects in the data. This aggregation of our analyses makes apparent that we encounter disagreements in variable identification between OLS and Lasso, especially in settings that are typical for everyday data analysis in biomedicine.

case of lung capacity prediction, the predictive variable selection concurred with the highest absolute coefficient in both approaches to determined importance. Here the prediction regime has probably missed the mechanistically relevant influence of smoking on lung capacity by pragmatic predictions based on body height alone. The high significance of all input variables may have been facilitated by the comparably high sample sizes.

DISCUSSION

Exploring a battery of empirical simulations and several biomedical datasets, our study offers insight into the possibly asymmetric tendencies between seeking accurate predictions in new individuals and identifying statistically significant effects across individuals.[27] Prediction and inference share a first step of estimating linear model coefficients; however, the difference arose in what we decided to do next with the estimated model effect sizes.

Charting a broad spectrum of data-analysis scenarios possible in everyday research, statistically significant relationships were not in all examined cases a guarantee to also achieve successful predictions when applying the model to other individuals. To restate, quantitative models varied between accounting for little or almost 100% variance in a target variable in fresh data points, even if a given model's effects were declared robust at the conventional significance level of $p < 0.05$. By contrast, effects not significant at $p < 0.05$ mostly failed to deliver useful predictions in data from unseen individuals. In short, our results are in line with the view that successful prediction in future or other data may often be more difficult to achieve than significance based on null-hypothesis testing. Of relevance for everyday research practice, predictability appears to be a demanding criterion. This is because, in our study, even small predictive performances typically coincided with finding underlying significant statistical relationships in the majority of the cases. However, even statistically strong associations with very low p values often shed only modest light on their value for the goal of prediction.

More broadly, researchers in most empirical sciences face questions of data analysis. What does it mean that a particular effect or model coefficient is "important" or not[28]? Statistical significance identified important coefficient effects based on (in-sample) deviation from a theoretical non-effect that is unlikely explained by noise.[29] Out-of-sample prediction, instead, discarded unimportant variables if the omission did not diminish the practically observed model performance on unseen data points. p values were computed by whether an effect size estimate (i.e., model coefficient) would take the actually obtained value at most 1 in 20 times if its impact on the outcome is not important.[29,30] In fact, an official report of the American Statistical Association (ASA) emphasized that "statistical significance is not equivalent to scientific, human, or economic significance."[12,31]

Concretely, in common-variant genetics, p values are typically derived from linear models to identify the most promising genetic locations among ~1,000,000 candidates, without relying on strong a priori knowledge.[11] Furthermore, an association between a candidate gene and diabetes grounded in a statistically significant p value may not necessarily imply that the same gene can always be used to successfully predict whether a given individual is affected by that disease. We used a predictive method that considered variable "importance" in a different way. A variable was considered relevant if leaving it out hurt the overall prediction accuracy when applying the previously built model was explicitly checked on fresh observations. Some authors believe that such empirical validation procedures to establish importance will be endorsed more in the future, and may increase due to adoption of code and data sharing. This ongoing evolution in data-analysis practices can promote across-study and across-method confirmation.[32]

Moreover, "importance" in quantitative research has probably no uniform theoretical basis.[1,28,33] The notion of variable relevance can therefore take different forms and shapes even in the family of the canonical linear model. A statistical method that produces importance assessments still requires the informed judgment of the investigator as to how far the

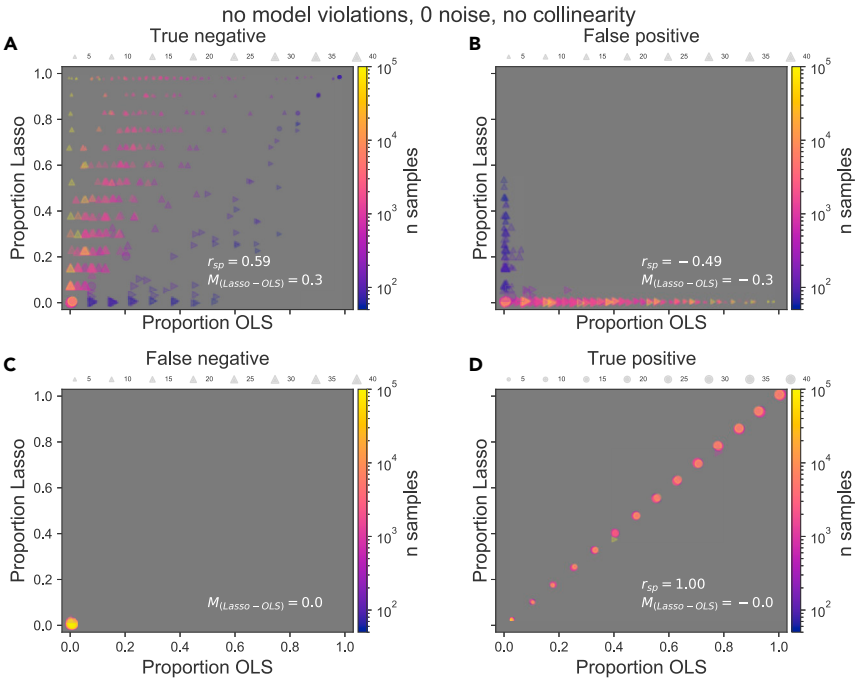

Figure 6. Well-Specified Modeling Setting: Divergence as a Function of Variable Selection Outcome

To detail the disagreements in variable identification between Lasso and OLS, we computed the proportion of encountered true negatives (A), false positives (B), false negatives (C), and true positives (D) for OLS (x axis) and Lasso (y axis). Circles indicate equal performance and triangles diverging performance. The size of the triangles and markers indicates the number of relevant variables to be recovered. To measure overall agreement or disagreement, we computed Spearman's rank correlation (r_{sp}) between the accuracies of Lasso and OLS as well as the mean of their differences $M_{(Lasso-OLS)}$. Jittering mitigated overplotting for display of simulation density. Across several thousand data simulations, these quantitative summaries revealed strong and systematic divergences between Lasso and OLS in the occurrence of true-negative hits and false-positive errors.

conclusions should be trusted. The initial choice of analysis method may be more or less well aligned with the substantive research question. Put differently, using p values or prediction accuracies for backing up research claims both have flaws and each is insufficient in some way.[2,26,34] Notably, we did not compute the usual p values on the automatically selected (non-zero) effect size estimates corresponding to input variables.[33] As one explanation, data-driven model selection would have undermined hypothesis-driven statistical inference. The initial prediction-based filtering step would have altered the sampling distribution of the variable coefficient estimates for subsequent significance-based variable filtering, which is an active area of research.[33,35] The ASA statement recommended that "no single index should substitute for scientific reasoning,"[12] a viewpoint shared by other prominent investigators.[36,37] In particular, Szucs and Ioannidis recently stressed monocultural training of biomedical scientists in statistical null-hypothesis testing as one reason behind frequent misuses of p-value-based methods, rather than increased reliance on effect sizes or out-of-sample prediction metrics.[30]

As a limitation of the present work, more extensive diagnostics could have shed light on how correlated versus non-correlated input variables behave in Lasso versus OLS in more detail. It is known that, embracing the prediction goal, the increased volatility of coefficient estimates of (multi-)collinear input variables does not typically change the explained variance of the outcome by the whole set of coefficients of the model.[25,38] However, confirming by the inference goal, such instability does affect the effect sizes and p values corresponding to individual model coefficients. Consequently, the coefficient of a single input variable can turn out to be not significant by itself, but become significant as part of a model with other input variables. Additionally, an actually significant effect can be subject to masking and disappear.[26] Future research is hence needed to better disentangle the distinct behavior of partial R^2 performances of the uncorrelated versus correlated subsets of model coefficients.

Comparing our quantitative investigation with related work, Lo and colleagues[39] stand out as targeting the identical aim of contrasting the relative merits of predictive and significant variables. Despite this similarity, Lo et al. primarily consider the question of when statistically significant variables are also found to be predictive or not. Our study instead did equally consider both directions: when significant variables are also predictive or not, as well as in which data-analysis scenarios predictive variables are also significant or not. Their study and ours concur in trying to expose similarity and differences of significance and predictability through concrete data-analysis scenarios. However, Lo and colleagues focus on what they call three "artificial examples" as thought experiments. Our study instead revisited four real-world biomedical datasets and analyzed >100,000 concrete data simulations covering a breadth of practical data-analysis scenarios that scientists and analysts are facing every day. Finally, the previous study concludes by "[encouraging] exploration away from significance-based methodologies and toward prediction-oriented ones."[39] In contrast, the core recommendation from our study is encouraging awareness for "horses for courses" rather than "if all you have is a hammer, everything looks like a nail."

Conclusion

Our quantitative investigation exposed how linear models—a workhorse in many areas of biomedical research—can be used with distinct and partly incompatible motivations. Using these analytical tools for the purpose of inference is useful for uncovering characteristics of biological processes. Using linear modeling for the alternative purpose of prediction is particularly suited for pragmatic forecasting of biological processes, potentially including clinical endpoints in individual patients.

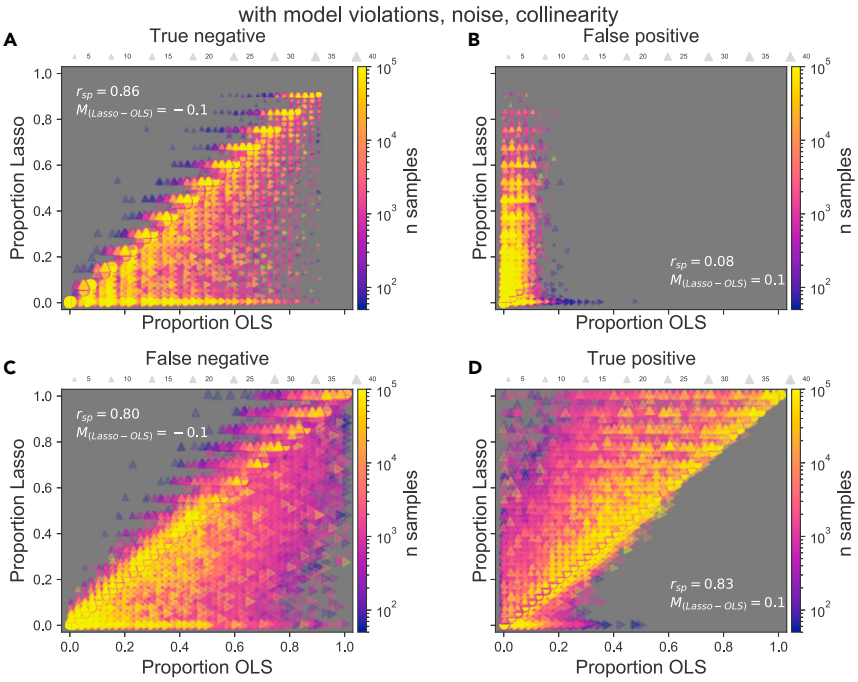

Figure 7. Ill-Specified Modeling Setting: Divergence as a Function of Variable Selection Outcome

To detail the disagreements in variable identification between Lasso and OLS, we computed the proportion of encountered true negatives (A), false positives (B), false negatives (C), and true positives (D) for OLS (x axis) and Lasso (y axis). Circles indicate equal performance and triangles diverging performance. The size of the triangles and markers indicates the number of relevant variables to be recovered. To measure overall agreement or disagreement, we computed Spearman's rank correlation (r_{sp}) between the accuracies of Lasso and OLS as well as the mean of their differences $M_{(Lasso-OLS)}$. For display of simulation density, agreement between OLS and Lasso would show a strong diagonal. Instead, the collective findings make apparent that systematic disagreements between Lasso and OLS emerge in all cases and were most pronounced for committed false-positive errors.

Some quantitative analysts therefore proposed that data-analysis applications should be primarily distinguished by the modeling goal, rather than strictly cataloging each method under a broad umbrella term, such as "statistics" versus "machine learning," "hypothesis-based" versus "data-driven," or "confirmatory" versus "exploratory."[3,32] It is critical for investigators and practicing medical doctors to acknowledge the partly incongruent modeling philosophies of drawing statistical inference and seeking algorithmic prediction, as well as their possibly non-identical scopes of interpretation.[1,27,6] Quantitative literacy may become increasingly relevant for taking rigorous and reproducible steps on our journey toward personalized medical care, with the prospect of benefiting the well-being of suffering patients.

More broadly, the prediction-inference dilemma may also recapitulate some ideas of Claude Bernard, a pioneer of controlled experiments in biomedicine.[40] Prediction may be closer to what he called "empirical medicine" oriented toward practical patient care as an often theory-free endeavor, such as symptom monitoring, risk assessment, and recommending therapeutic interventions. Statistical inference may bear a more direct relationship with his conceptualization of "scientific medicine" aimed at elucidating unknown principles underlying biological processes driven by theory. This is the case such as when asking for the reasons why certain individuals are at risk for disease onset or illuminating why a certain drug works better in some individuals than others.

In approaching a future of precision medicine, it may become central that modeling for inference and modeling for prediction are related but importantly different. The relevant subset of variables identified based on significant p values or based on predictive value can converge or diverge depending on the particular data scenario. We empirically demonstrated that diverging conclusions can emerge even when the data are the same and widespread linear models are used.[10] It must be noted that our message is not that investigators should never use predictive models when the goal of inference is desired. At this point, extracting inference from predictive models may involve an extra effort in many cases. Therefore, awareness of the relative strengths and weaknesses of both "data-analysis cultures" is unavoidable in fully benefiting from the accelerating data deluge in biology and medicine.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Associate Professor Danilo Bzdok, M.D., Ph.D. Email: danilo.bzdok@mcgill.ca.

Department of Biomedical Engineering, McConnell Brain Imaging Center (BIC), Montreal Neurological Institute (MNI), Faculty of Medicine, McGill University, Montreal, CanadaCanada CIFAR AI Chair, Mila – Quebec Artificial Intelligence Institute, Montreal, Canada. Email: bzdokdan@mila.quebec

Materials Availability

Readers wishing to explore in real time >100,000 simulation results on the relationship between inference and prediction can use our WebApp (binder-enabled): https://mybinder.org/v2/gh/banilo/inf_vs_pred_2018/r0_bioRxiv.

Data and Code Availability

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 OLS regression and corresponding p values (http://statsmodels.github.io). The *scikit-learn* package[41] provided efficient, unit-tested implementations for handling state-of-the-art machine-learning procedures (http://scikit-learn.org). All analysis scripts and necessary data that reproduce the results of the present study are readily accessible and open for reuse (https://github.com/banilo/inf_vs_pred_2020). The repository also provides extended Jupyter notebooks with additional analyses and an interactive WebApp.

Linear Model Analyses for Inference

To assess which variables have a statistically significant relation to the outcome, we evaluated the strength of evidence using OLS regression. Statistical significance was assessed by considering all candidate measures in the

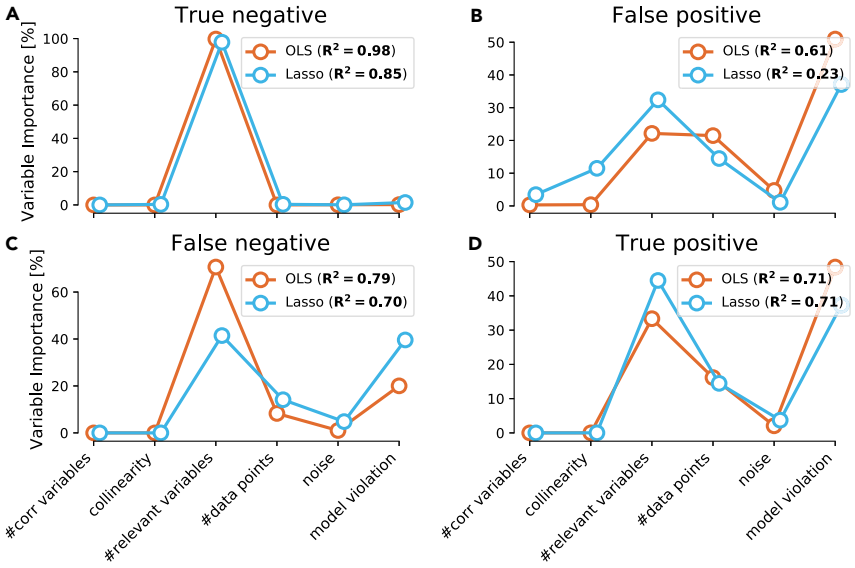

Figure 8. Driving Factors for Encountering Inference-Prediction Divergence across Dataset Simulations

Random Forest (RF) regression served as a "meta"-model to summarize error and performance behavior of OLS and Lasso as a function of our experimental factors: true negative (A), false positive (B), false negative (C), and true positive (D). This model-based description aggregated across the obtained simulation results to quantify the role of each particular factor (x axis) based on impurity reduction as a measure of impact (y axis) as well as the corresponding (out-of-bag) R^2 scores for each of the eight RF models summarizing the collection of results (every line represents one RF model). This high-level description across outcomes of data scenarios uncovered which experimental factors consistently drove differences in the recovery performance of OLS and Lasso. The driving experimental factor for our collective results turned out to be the number of truly relevant variables, which are commonly unknown in biomedical data analysis in practice.

same model.[34,42] We have performed least-squares regression to optimize the following objective:

$$min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_1\beta_1 - x_2\beta_2 - \ldots - x_p\beta_p)^2\right\},$$

where n is the number of individuals in the dataset, p is the number of input variables x_i (i.e., *independent*, *explanatory*, or *predictor variables*) for each individual, and y is the outcome measure (i.e., *dependent* or *explained variable*) that is to be expressed as a weighted sum of the variables x. Prior to model fitting, the data x were standardized by mean centering to zero and variance scaling to 1, across individuals. Given that all input variables were simultaneously present in the model, the approach responded to the question of relative contributions of each of the input variables in explaining the output y.

After model fitting has yielded a point estimate of each β coefficient, statistical inference was drawn to determine whether the contribution of input variable x_i in explaining the response y was sufficient to be deemed statistically significant (cf. below). The relevance of the effects is computed based on the confidence intervals of the β coefficients.[43] We drew inferential conclusions by formally testing for deviance of the observed effects from the null hypothesis. This approach attempted to reject the null hypothesis that the β coefficients were truly zero, that is, bear no coherent relation to the response variable y. Here, a non-significant β coefficient suggested that the variable at hand could be dropped from the model with little or no impact on explaining the output variable (which is, however, not explicitly evaluated). For these significance-based analyses, following typical applications of null-hypothesis testing, the p value was computed on the entire data from all considered individuals.

Linear Model Analyses for Prediction

For comparison with traditional linear regression, we chose an important recent modification of the linear model as a representative method for predictive pattern-learning algorithms.[44] Lasso estimated a weighted combination of the input variables, but the analysis goal revolved around prediction. We opted for this method because it is arguably the simplest existing method with an in-built sparsity constraint.[26] This additional assumption enforced that not all input variables were expected to be relevant in the final linear model solution (sparsity property of the Lasso). Instead, we have accommodated the biasing of model parameter estimation toward zero (shrinkage property of the Lasso, cf. below). This approach ensured that each variable had the same chance to be left out in the final model tuned for prediction in new observations.[26] We have thus identified subsets of the input variables that allowed for the strongest predictive effects. Automatic variable selection was achieved by minimizing the following optimization objective augmented with a penalty term during model estimation:

$$min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_1\beta_1 - x_2\beta_2 - \ldots - x_p\beta_p)^2 + \lambda\|\beta\|_1\right\},$$

where n is the number of individuals, p is the number of input variables x (i.e., *features*) measured for each individual, and y is the outcome to be predicted (i.e., *target variable*) by expressing it as a weighted sum of the standardized variables x. The hyper-parameter λ controlled the pressure for variable selection imposed during model fitting—the degree of the sparsity constraint. The higher the picked λ, the stronger was the tendency to set some coefficients β_i to exactly zero, which effectively "silenced" the corresponding measure's influence on the output variable. After model fitting, the model was applied to other data points to predict unobserved outputs (or, conceptually, "shipped" to other laboratories for repeated application). Within the cross-validation framework (Figure 1.), the selected model thus automatically chose the minimal subset of predictive variables that served to forecast the outcome.

Following model estimation, the practical performance of the candidate predictive model was evaluated based on standard cross-validation.[25] An explicit empirical measure of model performance was thus obtained to answer the question how much the predictive algorithm could be expected to generalize to data that would be seen in the future. To this end, model parameters were estimated on some data points, while the emerging model was explicitly put to the test in some independent, unseen data points.[45] To obtain this quantity of prediction accuracy, the linear model was first built on a larger part of the dataset. Second, emerging candidate models were evaluated and selected on unused data to avoid an overoptimistic evaluation of goodness-of-fit.[25] The out-of-sample prediction performance on the testing data samples thus quantified how likely the same pattern could be detected in future. In this way, the cross-validation scheme quantified the *out-of-sample performance* as an estimate of a model's capacity to generalize to data points acquired in the future. This assumption enforced that not all input variables were expected to be relevant in the final linear model solution (*sparsity* property of the Lasso). Instead, we have accommodated the biasing of model parameter estimation toward zero (*shrinkage* property of the Lasso, cf. below). This common modification (debiasing) allowed disentangling 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 linear regression (without shrinkage bias) based on the full set or subset of input variables automatically selected from the preceding Lasso estimation. Our approach to

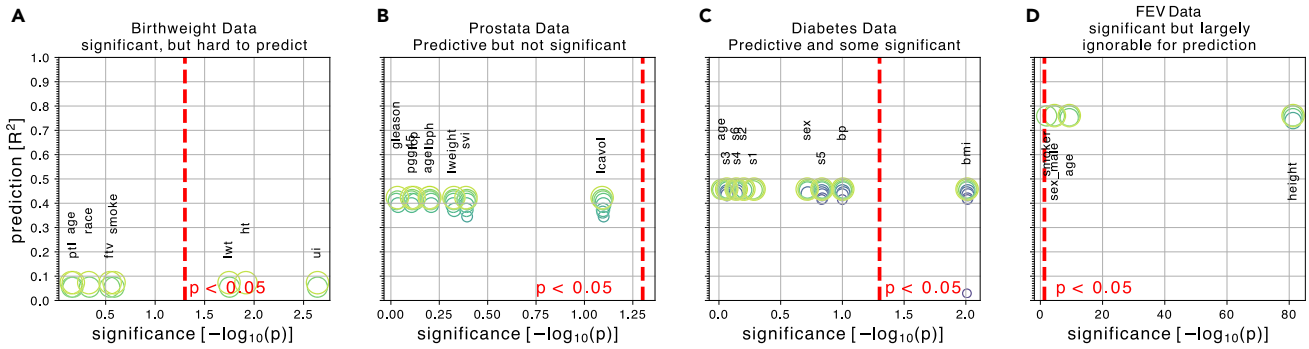

Figure 9. Predictability versus Significance in Four Medical Datasets

Integrative plots depict the inferential importance of each linear model coefficient (p values on x axis, log-transformed) and the predictive importance of coefficient sets (out-of-sample R^2 scores on y axis, obtained from model application on data not used for model fitting). Circles depict variable selection for candidate models with different regularization strengths (cf. Experimental Procedures).

(A) The body weight is to be derived from eight measures in 189 newborns. Three out of eight measures are statistically significantly associated with birth weight at $p < 0.05$ (red line). However, using the linear model for prediction explained only 8% of the variance in new babies ($R^2 = 0.08$).

(B) Prostate-specific antigen (PSA), a molecule for prostate carcinoma screening, is to be derived from eight measures in 87 men. None of the eight coefficients reached statistical significance based on common linear regression, although the fitted coefficients of the predictive model achieved 42% explained variance in unseen men.

(C) Disease progression after 1 year is to be derived from 10 measures in 442 diabetes 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 achieve 46% explained variance in independent patients.

(D) Lung capacity as quantified by forced expiratory volume (FEV) is to be derived from four measures in 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 four coefficients ($R^2 = 0.74$ versus $R^2 = 0.76$).

To expose the trade-off between parsimony and prediction performance, the circles (green) show the sets of Lasso coefficients at different sparsity levels. In sum, linear models can show all combinations of predictive versus not and significant versus not in biomedical data analysis.

modeling for prediction was centered around evaluating the capacity of already extracted models to derive quantities of interest from new data points, potentially later encountered individuals. This form of building models from data has been explicitly optimized for and was naturally applicable to a single data point or person.

From a more general theoretical perspective, as the hyper-parameter λ gets close to zero, the asymptotic properties of Lasso approach those of ordinary linear regression.[26,46] At a small λ, Lasso converged to a full set of non-zero coefficients (no variable selection, analogous to ordinary linear regression), which are, however, skewed toward zero (active shrinkage, unlike ordinarily linear regression). In addition to this "estimation bias," increasing the tuning parameter λ further, Lasso converges to skewed coefficients, an increasing subset of which start to show exactly zero coefficients. If we knew that the observed data were generated from a sparse loading vector, would the Lasso actually recover the true sparsity pattern when the number of observed data points grows (defined as asymptotic consistency)? Previous formal investigations showed[26,46] that, the higher the noise in the data or the more collinearity between the input variables, the more Lasso can show erratic behavior, as this estimator is especially vulnerable to converge to solutions with more mistakes in variable selection. Furthermore, the more input variables at play, the more inducing shrinkage and variable selection by Lasso's penalty term will converge to a model solution that can be expected to outperform ordinary linear regression in terms of prediction accuracy in new data points. On the other hand, the number of relevant variables that Lasso can automatically select is bounded by the sample size.[26,46] Thus, the final solution can be always less sparse, that is, contain more active (non-zero) coefficients, as the number of data points used for model fitting keeps growing. Moreover, the Lasso is known not to be consistent in the presence of strong intervariable correlation with unlimited data points. Yet there is a lack of comprehensive practical benchmarks that delineate the consequences for empirical data analysis with finite sample size.

Empirical Data Simulations

Our synthetic data-simulation approach was motivated by the realization that formal guarantees for the expected model prediction performance have often been challenging to derive by mathematical theory for a particular number of available data points.[10,45] Instead, we opted for empirical simulations as an alternative access route for studying the properties of the statistical methods in computational experiments.[43] In this way, we have directly confronted linear modeling for inference and for prediction in a series of synthesized datasets, columns of input variables X, each related or not related to the outcome y. Each dataset was generated from a set-up ground-truth model $y = X\beta + \epsilon$, where β are fixed random coefficients, X is a data matrix containing n data points and p variables with random entries drawn from a standard Gaussian distribution $N_{(\mu=0,\sigma=1)}$, and ε denotes the added Gaussian noise. Each dataset was fed into the linear models with the aim of identifying significant input measures or identifying input measures most useful for accurate predictions on new observations (cf. above).

To sharpen the distinction between explanatory and predictive modeling in general, we have systematically varied distinct aspects of the data-generating process as follows.

(1) Samples-to-variables ratio. To investigate the relation between the number of data points n relative to the number of variables p, we gradually scaled the number of available observations. We covered the lower range between 50 and 100 data points in steps of 10, which probably well represents a majority of studies in biomedicine. Between 100 and 2,000 data points, we increased the sample size in steps of 100. Moreover, we considered the extreme cases 10,000 and 100,000 data points, which acknowledges 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.

(2) Proportion of informative variables. To study how the fraction of relevant versus irrelevant variables modulates the inferential and predictive processes, we have varied the proportion of non-zero β coefficients in the ground-truth model used for generating X. We considered 14 proportions ranging from only 1 to all 40 input variables carrying information about the response y. Note that this breadth of experimental scenarios covers the number of considered variables in the analyses of real-world biomedical datasets (cf. below).

(3) Redundant versus unique sources of information: To elucidate how correlated input measures trade off against each other with respect to the outcome, we have introduced different degrees of pairwise covariation between the variable columns of X (i.e., collinearity). Ground-truth models also generated data from a multivariate Gaussian distribution that exposed 50% or 90% of common variation between the relevant variables, complementing datasets that contain only mutually independent variables (i.e., 0% covariation).

(4) Signal-to-noise (SNR) ratio. To assess the role of nuisance variation in the data, such as induced by imperfect measurement techniques, we have systematically manipulated the noise ε in how the real model relates to the response y. The nuisance term—as well as the input data (cf. above)—was generated from $N_{(\mu=0,\sigma=1)}$ and multiplied by 0.5, 1, 2, 5, 10, or 0 (i.e., generating data without any noise). Given the definition of SNR as Var(true mean)/Var(noise), our simulations of observed datasets covered a range of SNR scenarios that are practically realistic.

(5) Model violations. Finally, we examined how inference and prediction behave when the linear model cannot fully capture how the data came about. We have therefore introduced alterations on 50% of the relevant variables in X that lead to associations between the input variables and the outcome inconsistent with the linear model specification. In addition to datasets with exclusively linear effects (i.e., we can find the true model), deviations between the generating and fitting model were introduced by one of several data transformations: taking the absolute value, the natural logarithm, the exponential, the square root, the multiplicative inverse, or polynomial expansion of degree 2–5.

The collection of simulated datasets has realized 113,400 different data-analysis scenarios. For each case, we focused on the best (smallest) p value among all input variables in the model and the highest prediction performance of the overall model as quantified by the (out-of-sample) R^2 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 analyses required roughly 4 weeks of computation time and produced 2 Gb of modeling results.

ACKNOWLEDGMENTS

We are indebted to Olivier Grisel and Gael Varoquaux for fruitful discussion on the topic (both INRIA Saclay/France). We thank several investigators for insightful comments on a previous version of the manuscript: B.T. Thomas Yeo (National University of Singapore), Guillaume Dumas (Institut Pasteur/France), Nikolaus Kriegeskorte (Columbia University/USA), Daniele Marinazzo (Ghent University/Belgium), Benjamin de Haas (University of Giessen/Germany), and João Sato (Universidade Federal do ABC/Brazil).

D.B. was supported by National Institutes of Health (NIH R01 AG068563A), the Canadian Institutes of Health Research (CIHR 438531), the Healthy Brains Healthy Lives initiative (Canada First Research Excellence fund), Google (Research/Teaching Award), and by the CIFAR Artificial Intelligence Chairs program (Canada Institute for Advanced Research). D.B. was also 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, and 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 research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project). D.E. acknowledges support by the Amazon AWS Research Grant (2015) and the German National Merit Foundation, as well as the French National Institute for Informatics and Automation (INRIA) (Starting Researcher Position SRP 2016).

AUTHOR CONTRIBUTIONS

D.B. and B.T. conceived the project. D.B. and D.E. performed the research and wrote the manuscript. B.T. provided analytical solutions and provided feedback on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 16, 2020
Revised: August 17, 2020
Accepted: September 14, 2020
Published: October 8, 2020

REFERENCES

1. Breiman, L. (2001). Statistical modeling: the two cultures. Stat. Sci. *16*, 199–231.
2. Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. Nat. Methods *15*, 233–234.
3. Bzdok, D., and Ioannidis, J.P.A. (2019). Exploration, inference and prediction in neuroscience and biomedicine. Trends Neurosci. *42*, 251–262.
4. Bzdok, D., Nichols, T.E., and Smith, S.M. (2019). Towards algorithmic analytics for large-scale datasets. Nat. Machine Intelligence *1*, 296–306.
5. Bzdok, D., Varoquaux, G., and Steyerberg, E.W. (2020). Prediction, not association, paves the road to precision medicine. JAMA Psychiatry. https://doi.org/10.1001/jamapsychiatry.2020.2549.
6. Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. Front. Neurosci. *11*, 543.
7. Cox, D.R. (2006). Principles of Statistical Inference (Cambridge University Press).
8. Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues, G. Keren and C. Lewis, eds. (Lawrence Erlbaum Associates), pp. 311–339.
9. Efron, B., and Tibshirani, R.J. (1991). Statistical data analysis in the computer age. Science *253*, 390–395.
10. Efron, B., and Hastie, T. (2016). Computer-Age Statistical Inference (Cambridge University Press).
11. Efron, B. (2012). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Cambridge University Press).
12. Wasserstein, R.L., and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. Am. Stat. *70*, 129–133.
13. Ioannidis, J.P. (2018). The proposal to lower p value thresholds to. 005. JAMA, 1429–1430.
14. Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017). The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ *5*, e3544.
15. Blei, D.M., and Smyth, P. (2017). Science and data science. Proc. Natl. Acad. Sci. U S A *114*, 8689–8692.
16. Leonelli, S. (2016). Data-Centric Biology: A Philosophical Study (University of Chicago Press).
17. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. Technical Report (McKinsey Global Institute). https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.
18. Goodfellow, I.J., Bengio, Y., and Courville, A. (2016). Deep Learning (MIT Press).
19. Shmueli, G. (2010). To explain or to predict? Stat. Sci. *25*, 289–310.
20. Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507.
21. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., and Webster, D.R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. *2*, 158–164.
22. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., and Ng, A.Y. (2019). Cardiologist-level arrhythmia

detection and classification in ambulatory electrocardiograms using a deep neural network. Nat. Med. *25*, 65–69.

23. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.
24. Carr, D.B., Littlefield, R.J., Nicholson, W., and Littlefield, J. (1987). Scatterplot matrix techniques for large N. J. Am. Stat. Assoc. *82*, 424–436.
25. Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning (Springer Series in Statistics).
26. Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical Learning with Sparsity: The Lasso and Generalizations (CRC Press).
27. Woo, C.-W., Chang, L.J., Lindquist, M.A., and Wager, T.D. (2017). Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. *20*, 365–377.
28. Thompson, B., and Borrello, G.M. (1985). The importance of structure coefficients in regression research. Educ. Psychol. Meas. *45*, 203–209.
29. Casella, G., and Berger, R.L. (2002). Statistical Inference (Duxbury).
30. Szucs, D., and Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. Front. Hum. Neurosci. *11*, 390.
31. Matthews, R., Wasserstein, R., and Spiegelhalter, D. (2017). The ASA's p-value statement, one year on. Significance *14*, 38–41.
32. Donoho, D. (2017). 50 years of data science. J. Comput. Graph. Stat. *26*, 745–766.
33. Taylor, J., and Tibshirani, R.J. (2015). Statistical learning and selective inference. Proc. Natl. Acad. Sci. U S A *112*, 7629–7634.
34. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics *25*, 714–721.
35. Zhang, C.H., and Zhang, S.S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Series B Stat. Methodol. *76*, 217–242.
36. Cohen, J. (1990). Things I have learned (so far). Am. Psychol. *45*, 1304.
37. Gigerenzer, G., and Murray, D.J. (1987). Cognition as Intuitive Statistics (Psychology Press).
38. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning (Springer).
39. Lo, A., Chernoff, H., Zheng, T., and Lo, S.H. (2015). Why significant variables aren't automatically good predictors. Proc. Natl. Acad. Sci. U S A *112*, 13892–13897.
40. Bernard, C. (1957). An Introduction to the Study of Experimental Medicine (Courier Corporation).
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
42. Freedman, D.A. (1983). A note on screening regression equations. Am. Stat. *37*, 152–155.
43. Gelman, A., and Hill, J. (2007). Data Analysis Using Regression and Multilevel Hierarchical Models (Cambridge University Press).
44. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. *58*, 267–288.
45. Shalev-Shwartz, S., and Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press).
46. Bühlmann, P., and Van De Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications (Springer Science & Business Media).
47. Jordan, M.I., and Mitchell, T.M. (2015). Machine learning: trends, perspectives, and prospects. Science *349*, 255–260.